NEW SUPPORT VECTOR MACHINE FORMULATIONS AND ALGORITHMS WITH APPLICATION TO BIOMEDICAL DATA ANALYSIS

A Thesis
Presented to
The Academic Faculty

by

Wei Guan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing

Georgia Institute of Technology
August 2011

NEW SUPPORT VECTOR MACHINE FORMULATIONS AND ALGORITHMS WITH APPLICATION TO BIOMEDICAL DATA ANALYSIS

Approved by:

Professor Alexander Gray, Advisor
College of Computing
Georgia Institute of Technology

Professor Shamkant Navathe
College of Computing
Georgia Institute of Technology

Professor John McDonald
Department of Biology
Georgia Institute of Technology

Professor Charles IsBell
College of Computing
Georgia Institute of Technology

Professor Jeffrey Skolnick
Department of Biology
Georgia Institute of Technology

Date Approved: May 9, 2011

ACKNOWLEDGEMENTS

This thesis would not have been possible without the generous support from many people inside and outside of the College of Computing at Georgia Tech. I would like to express my gratitude to all of them.

First and foremost, my deepest gratitude is to my advisor, Professor Alexander Gray, who provided me tremendous support, advice and guidance for my study, research, and career development. Not only has he imparted to me a wealth of knowledge in machine learning and computational data analysis, but also he has taught me the skills essential to a PhD student and a researcher, including the general problem-solving skills, system-level thinking and the ability to identify and define research problems. I have been fortune to study under his tutelage.

I would like to thank my thesis committee members Professor Shamkant Navathe, Professor John McDonald, Professor Jeffrey Skolnick, and Professor Charles Isbell for their constructive feedbacks of my work that have greatly improved the quality of this thesis. Special thanks to Professor Facundo Fernández from the School of Chemistry and Biochemistry and Professor Mark Borodovsky from School of Computer Science and Engineering. Their challenging questions and bountiful support contributed to the success of my research projects.

I would also like to thank the Data Mining Research group of Yahoo, the Medical Text and Image Analysis group of IBM Research, especially my mentors, Dr. Pavel Berkhin, Dr. Long-ji Lin, Dr. Christina Yip Chung, Dr. Anni Coden, Dr. Igor Sominsky and Dr. Michael Tanenblatt for offering me interesting, valuable internship opportunities. The experience in these industrial research labs has broadened my prospective and helped to develop my research skills.

I would like to thank all of my friends and colleagues at Georgia Tech, Dr. Nathan Bowen, Dr. Christina Hampton, Dr. Manshui Zhou, Dr. Arkadas Ozakin, Dr. Minh Quoc Nguyen, Saurav Sahay, and Qinyi Wu, just to mention a few.

Finally, I wish to thank my parents, my brother and my sister-in-law for their support and encouragement through all the years. To them, I dedicate this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The Support Vector Machine (SVM) classifier seeks to find the separating hyperplane $wx = r$ that maximizes the margin distance $1/||w||_2^2$. It can be formalized as an optimization problem that minimizes the hinge loss $\sum_i (1 - y_i f(x_i))_+$ plus the L_2-norm of the weight vector. SVM is now a mainstay method of machine learning. The goal of this dissertation work is to solve different biomedical data analysis problems efficiently using extensions of SVM, in which we augment the standard SVM formulation based on the application requirements. The biomedical applications we explore in this thesis include: cancer diagnosis, biomarker discovery, and energy function learning for protein structure prediction.

Ovarian cancer diagnosis is problematic because the disease is typically asymptomatic especially at early stages of progression and/or recurrence. We investigate a sample set consisting of 44 women diagnosed with serous papillary ovarian cancer and 50 healthy women or women with benign conditions. We profile the relative metabolite levels in the patient sera using a high throughput ambient ionization mass spectrometry technique, Direct Analysis in Real Time (DART). We then reduce the diagnostic classification on these metabolic profiles into a functional classification problem and solve it with functional Support Vector Machine (fSVM) method. The assay distinguished between the cancer and control groups with an unprecedented 99% accuracy (100% sensitivity, 98% specificity) under leave-one-out-cross-validation. This approach has significant clinical potential as a cancer diagnostic tool.

High throughput technologies provide simultaneous evaluation of thousands of potential biomarkers to distinguish different patient groups. In order to assist biomarker discovery from these low sample size high dimensional cancer data, we first explore

a convex relaxation of the L_0-SVM problem and solve it using mixed-integer programming techniques. We further propose a more efficient L_0-SVM approximation, fractional norm SVM, by replacing the L_2-penalty with L_q-penalty (q in (0,1)) in the optimization formulation. We solve it through Difference of Convex functions (DC) programming technique. Empirical studies on the synthetic data sets as well as the real-world biomedical data sets support the effectiveness of our proposed L_0-SVM approximation methods over other commonly-used sparse SVM methods such as the L_1-SVM method.

A critical open problem in *ab initio* protein folding is protein energy function design. We reduce the problem of learning energy function for *ab initio* folding to a standard machine learning problem, learning-to-rank. Based on the application requirements, we constrain the reduced ranking problem with non-negative weights and develop two efficient algorithms for non-negativity constrained SVM optimization. We conduct the empirical study on an energy data set for random conformations of 171 proteins that falls into the *ab initio* folding class. We compare our approach with the optimization approach used in protein structure prediction tool, TASSER. Numerical results indicate that our approach was able to learn energy functions with improved rank statistics (evaluated by pairwise agreement) as well as improved correlation between the total energy and structural dissimilarity.

CHAPTER I

INTRODUCTION

One of the main focuses in Bioinformatics is to develop efficient tools and methods that are capable of analyzing and transforming the highly heterogeneous experimental data into biological knowledge about the underlying mechanism. However, the exponential growth of the amount of biological data available present great challenges in extraction useful information from these data. Therefore, machine learning techniques, for example supervised learning methods, have been widely used to assist researches in bioinformatics.

In this dissertation work, we extend support vector machine optimization, the current mainstay of machine learning, to efficiently solve the reduced machine learning problems from biomedical applications. To examplify our approach, we analyze problems from three different biological applications including cancer diagnosis, biomarker discovery and protein energy function learning. We investigate the coupling of a high throughput ambient ionization technique for mass spectrometry (MS) with machine learning approaches for the metabolomic classification of sera from ovarian cancer and control patients. We explore more aggressive feature selecting support vector machine methods for biomarker discovery from high throughput cancer data sets. We also study the problem of learning energy function for *ab initio* protein folding through machine learning approaches.

Overall, we believe that these new support vector machine formulations and algorithms could be the potential solutions for developing efficient machine learning tools for researches in life science domains.

1.1 Support Vector Machine Optimization

Given a dataset $S = \{x_i, y_i\}_{i=1}^m$ ($x_i \in R^n$ is the feature vector of ith data point and $y_i \in \{0, 1\}$ is the corresponding label), for two-class classification problems, support vector machine (SVM) learns the separating hyperplane $wx = \gamma$ that maximizes the margin distance $\frac{2}{||w||_2^2}$, where w is the weight vector and γ is the bias (see Figure 1).

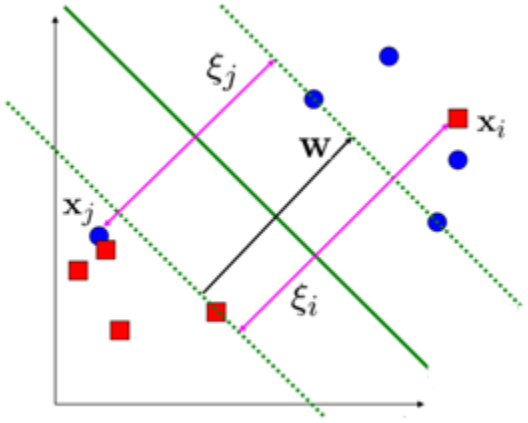


Figure 1: Illustration of Max Margin Classification

Defining ξ as the slack variables, $c > 0$ as the error penalty parameter, diagonal matrix $Y \in R^{m \times m}$ with $Y_{ii} = y_i$, data matrix $X = [x_1, x_2, \cdots, x_m]^T$, vector $e_k = [1, 1, \cdots, 1]^T \in R^k$, and identity matrix $I_k \in R^{k \times k}$, we can formulate the linear SVM learning problem into the following convex optimization problem.

$$\begin{aligned} \min \quad & \tfrac{1}{2} \mid\mid w \mid\mid_2^2 + c \mid\mid \xi \mid\mid_1 \\ \text{s.t.} \quad & Y(Xw - \gamma e_m) + \xi \geq e_m \\ & \xi \geq 0 \end{aligned} \tag{1}$$

Defining $\alpha \in R^m$ as the Lagrange multiplier and H as the kernel matrix with $H_{ij} = y_i y_j x_i \cdot x_j$, then the dual problem can be represented as:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\alpha^T H\alpha - e_m^T\alpha \\ & e_m^T Y\alpha = 0 \\ \text{s.t.} \quad & \alpha \leq ce_m \\ & \alpha \geq 0 \end{aligned} \tag{2}$$

The optimal weight vector is then computed as $w = \sum_{i=1}^{m} y_i\alpha_i y_i x_i$. α is usually a sparse vector as only support vector $x_k \in SV$ has non zero α_k values. The optimal decision function for a data point x is defined as $f(x) = w \cdot x - \gamma$. The prediction label is $+1$ if $f(x) > 0$ and -1 otherwise.

Because of its theoretical guarantees and superior empirical performance, SVM has become the interest of research in classification for decades and been successfully applied to various scientific problems including biomedical applications.

1.2 Examples of Biological Applications

1.2.1 Ovarian Cancer Diagnosis

Despite decades of research and an annual investment in the U.S. of more than $2 billion on treatment, ovarian cancer remains the leading cause of deaths from gynecological malignancies [16]. It is estimated that 21,880 women will be diagnosed with and 13,850 women will die of cancer of the ovary in 2010 [79]. Due to the asymptomatic nature of the disease, women are frequently undiagnosed until the disease is late in its progression (stage III/IV) when the 5-year survival rate is only 15-20% [80]. The assay for CA125 is currently the only FDA-approved test for ovarian cancer detection, however the overall positive predictive value of CA125 has been reported to be less than 10% [83].

Most ovarian cancer biomarker discovery studies are based on the univariate or multivariate anaysis of high throughput data focusing on qualitative or quantitative changes (e.g. methylation, glycosylation) of large biopolymers (e.g. DNA, RNA, glycans and proteins) [114]. In contrast, metabolic biomarker discovery approaches

that focus on small molecules (below 1 kDa) have received significantly less attention.

We investigate the coupling of a high throughput ambient ionization technique for mass spectrometry (MS), Direct Analysis in Real Time (DART), with machine learning approaches for the metabolomic classification of sera from ovarian cancer and control patients [126]. We reduce the task of classifying DART MS profiles into a functional classification problem and solve it using functional SVM method. By combining the DART-TOF MS with a customized fSVM classification algorithm, we were able to distinguish sera from cancer patients and controls with 99% accuracy using the stringent leave-one-out cross-validation evaluation (100% sensitivity and 98% specificity). We also estimate a clinically significant 12% PPV for our assay in this sub-population, which is above the minimum value (10%) for a test to be considered of clinical significance. We view this as a successful step towards the development of an accurate new approach to the diagnosis of ovarian and other cancers.

1.2.2 Biomarker Discovery

Biomarker discovery, also referred as gene selection, metabolite panel selection, aims to identify molecules (e.g., genes, metabolites) that are disease related. Biomarker discovery allows to make the diagnostic process cheaper and targeted, and to narrow down the number of biomarkers to better understand their biological significance. The task is typically reduced into classification-driven feature selection and subsequent biological validation.

High throughput technologies such as microarray, mass spectrometry, etc., provide simultaneous evaluation of thousands of potential biomarkers that distinguish different patient groups. Cancer data sets generated by high throughput technologies are usually low sample size, consists of only a few hundreds patients. Our goal is to develop more aggressive feature selection with similar or better accuracy than previous techniques on these low-sample size high-dimensional cancer data sets.

For this, we investigate two convex relaxations of the L_0-SVM formulation and propose efficient solution to the resulting optimization problems [43, 41]. Empirical study on both simulations and real-world data sets support the effectiveness of the fractional-norm SVM over other commonly-used sparse SVM methods. We believe our approach is a promising feature selection method for biomarker discovery form the low sample size high-dimensional data sets.

1.2.3 Protein Structure Prediction

Proteins are polymers assembled from 20 naturally occurring amino acids, which fold to a unique, biologically active, three-dimensional conformation called the *native structure*. Their functions are governed by their three-dimensional structures, which in turn are fully determined by their amino acid sequences. Predicting the native structure of a protein from its amino acid sequence, is one of the most important and challenging scientific problems in contemporary biology and chemistry [36], [100]. The experimental determination of protein tertiary structure is a time consuming and expensive process. Hence, computational methods play an essential role in the native structure prediction of proteins.

There are three classes of computational based protein structure prediction approaches: homology modeling, threading and *ab initio* folding. For a query protein, if none of its homologous sequences has an experimentally resolved structure, the only remaining approach to predict its native structure is *ab initio* folding.

Ab initio folding attempts to find the native structure of a protein "from scratch" (see Figure 12). A critical open problem in *ab initio* protein folding is protein energy function design, which pertains to defining the energy of protein conformations in way that makes folding most efficient and reliable. We address this issue as a weight optimization problem and demonstrate a machine learning approach, learning to rank, to

solve this problem. To maintain the physicality of the results, we constrain the problem with non-negative weights. We develop efficient algorithm to solve the resulting non-negative RankingSVM problem [40]. We demonstrate an energy function which maintains the correct ordering with respect to structure dissimilarity to the native state more often, is more efficient and reliable for learning on large protein sets, and is qualitatively superior to the current state-of-the-art energy function.

1.3 Thesis Organization

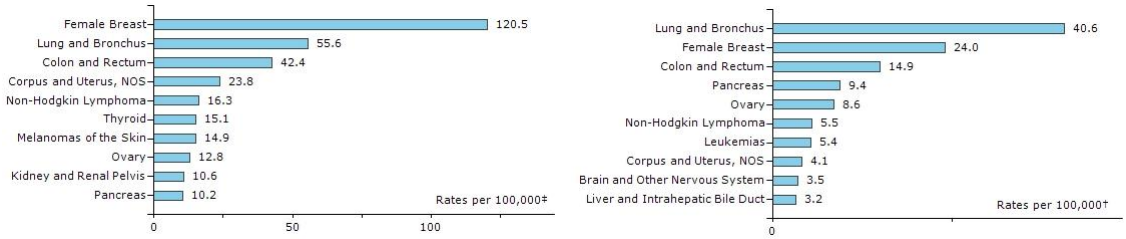
The rest of the thesis is organized as follows: In the next chapter, we investigate the coupling of a high throughput ambient ionization technique for mass spectrometry (MS) with functional support vector machine method for the metabolomic classification of sera from ovarian cancer and control patients. In Chapter 3, we propose two feature selecting support vector machine based convex relaxations of the L_0-SVM formulation to assist the task of biomarker discovery in low sample size and high dimensional cancer data sets. To better address the scalability issue on the low sample size high dimensional cancer data sets, we further propose the fractional-norm SVM problem and solve it through difference of convex programming technique. We reduce the task of learning energy function for *ab initio* protein folding into a machine learning problem, ranking-via-classification. The reduction, optimization methods and the results analysis are presented in Chapter 4. Conclusion and directions for future work are discussed in Section 5.

CHAPTER II

OVARIAN CANCER DETECTION FROM MASS SPECTROMETRIC METABOLIC PROFILING

2.1 Introduction

Ovarian cancer (OC) is the most lethal of gynecological cancers and the 5th leading cause of all cancer-related deaths among women [23] (see the statistics summary in Figure 2). It is estimated that 21,880 women will be diagnosed with and 13,850 women will die of cancer of the ovary in 2010 [79]. While the 5-year survival rate for women diagnosed with the disease early in its progression is greater than 90%, the survival rate for patients diagnosed at later stages is only 20% [54]. The main challenge with ovarian cancer is that it typically arises and progresses initially without well-defined clinical symptoms [57]. Thus, successful diagnosis plays a central role in deciding appropriate therapy and improving patient prognosis.



(a) Incident Rate (b) Death Rate

Figure 2: Top 10 Cancer Sites: 2003-2007, Female, United States

Although screening for specific biomarkers that are diagnostic of ovarian cancer has been an active area of research since the early 1970's [77], no effective diagnostic tests are yet available. The assay for CA125 is currently the only FDA-approved test for ovarian cancer detection but the overall predictive value of CA125 has been reported to be less than 10% [83]. Most ovarian cancer biomarker discovery studies are

based on the univariate or multivariate comparison of high throughput data focusing on qualitative or quantitative changes (e.g. methylation, glycosylation) of large biopolymers (e.g. DNA, RNA, glycans and proteins) [114]. In contrast, metabolic biomarker discovery approaches that focus on small molecules (below 1 kDa) have received significantly less attention, despite the fact that metabolic profiling of human serum has long been touted as a promising technology for the early detection of many diseases, including cancer [80]. In this trend, a few studies have reported individual metabolites potentially useful for ovarian cancer detection, the most studied being lysophosphatidic acid [7, 102, 117] and lipid associated sialic acid [84, 94, 95, 104, 110].

Since metabolites have vastly-differing chemical properties and occur in a wide range of concentrations, mass spectrometry (MS) is a preferred method for broadband metabolic profiling [33]. Although MS has been successfully applied in the development of proteomic biomarker panels using surface-enhanced laser/desorption ionization (SELDI) MS [83, 28, 71, 116] and matrix-assisted laser desorption/ionization (MALDI) MS [92, 1], technologies such as liquid chromatography (LC) MS for the effective analysis of the metabolome are still evolving as bioinformatic techniques for the analysis of the resulting cancer data [107],[44].

We investigate the coupling of a high throughput ambient ionization technique for mass spectrometry (MS) with machine learning approaches for the metabolomic classification of sera from ovarian cancer and control patients. This technique, known as Direct Analysis in Real Time (DART) [27], is one of the members of the rapidly growing family of open-air (ambient) ionization methods for MS [49] that also includes Desorption Electrospray Ionization (DESI) [81]. In this DART MS test, a stream of excited metastables is used to desorb and chemically ionize a dried drop of derivatized serum. A typical DART MS profile displays a multitude of signals corresponding to metabolites rapidly desorbed and ionized in a time-dependent fashion (Figure 3(c.x)). The classification of serum sample are further reduced into functional classification

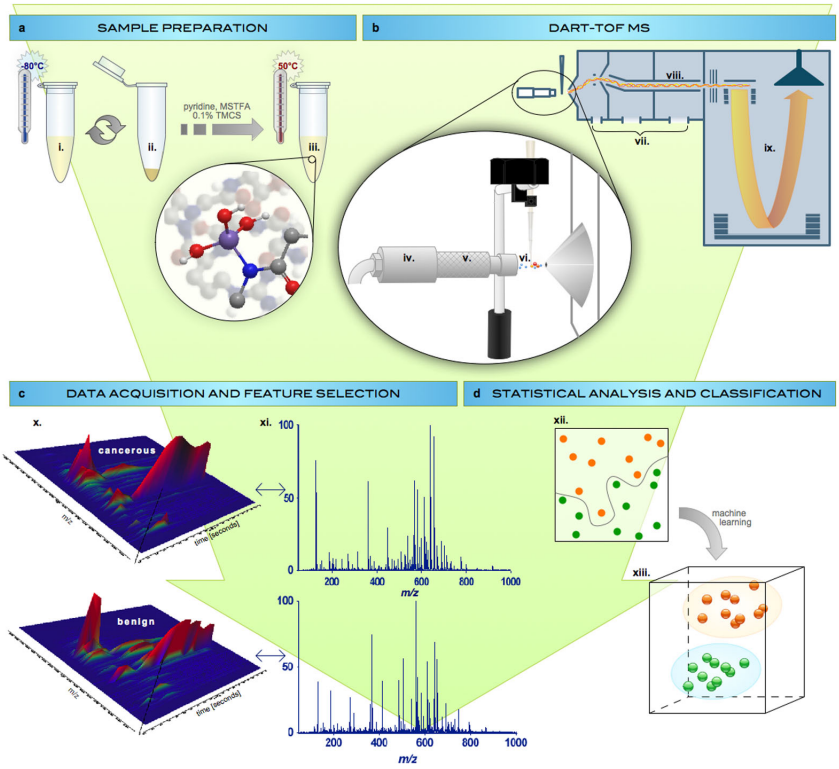
on these DART MS profiles, and solved with a customized functional Support Vector Machine (fSVM) algorithm. The assay distinguished between the cancer and control groups with 98.9% accuracy (100% sensitivity; 98-100% specificity) under leave-one-out cross-validation evaluation.

The rest of this chapter is organized as follows: In the next section, we briefly describe the workflow of the metabolomic investigation of patient blood serum samples by DART-TOF MS as illustrated in Figure 3. In section 3, we present the reduced functional classification problem on the DART MS profiles as well as the prediction performance evaluation on the cancer data using functional SVM method. In Section 4, we discuss the biological significance of our data analysis results. The results demonstrate the utility of this approach to derive panels of metabolic spectral features that are potentially useful for the diagnosis of ovarian cancer.

2.2 DART MS Cancer Data Preparation

2.2.1 Sample Collection

Serum samples were obtained from the Ovarian Cancer Institute laboratory at Georgia Tech after approval by the Institutional Review Board from Northside Hospital and Georgia Institute of Technology (Atlanta, GA; Table 1). All donors were required to fast and to avoid medicine and alcohol for $12h$ prior to sampling, except for certain allowable medications, for instance, diabetics were allowed insulin. Following informed consent by donors, $5m$L of whole blood were collected by venipuncture into evacuated blood collection tubes that contained no anticoagulant (blood taken prior to the administration of anaesthesia, immediately preceding surgery). Within one hour of venipuncture, serum was collected and 200μL aliquots of each sample were stored in $1.5m$L micro-tubes at -80°C until ready to use.



(a) Serum sample preparation: i. protein precipitation, centrifugation and separation of the metabolite containing supernatant followed by ii. evaporation of solvent to generate a metabolite-containing pellet. This pellet is then subject to iii. derivatization to increase volatility of polar metabolites. (b) Schematic of the DART-TOF mass spectrometer equipped with a custom-built sample arm (iv. glow discharge compartment, v. gas heater, vi. ionization region where sample-carrying capillary is placed), vii. differentially-pumped atmospheric pressure interface to transport ions towards the mass analyzer, viii. radiofrequency ion guide where ions are collisionally cooled prior to entering the ix. orthogonal TOF mass analyzer. (c) Typical data are acquired in a time-resolved fashion (x. three-dimensional contour plots of single runs corresponding to an ovarian cancer patient (top), and a control (bottom)). The region of the time-resolved signal with best signal-to-noise ratio was averaged yielding xi. profile mass spectra reflecting metabolic fingerprints. (d) Machine learning techniques such as SVMs are used for building a multivariate classifier (xii. objects in original variable space, xiii. objects in classifier space).

Figure 3: Diagram: Metabolomic Investigation of Serum Samples for Detection of Ovarian Cancer

Table 1: Patients analyzed in this study

Patient ID	Ovarian Histopathology	Stage/Grade	Age at Surgery
242	papillary serous carcinoma	IIIc/3	63
281	papillary serous carcinoma	III/1	66
454	papillary serous carcinoma	III/3	72
458	papillary serous carcinoma	IIIc/3	59
472	papillary serous carcinoma	IIIc/2-3	49
473	papillary serous carcinoma	IIIc/3	48
491	papillary serous carcinoma	I/	74
495	papillary serous carcinoma	IIIb/3	43
512	papillary serous carcinoma	IIIb/2-3	59
517	papillary serous carcinoma	Ia/3	59
526	papillary serous carcinoma	IIIc/2-3	49
528	papillary serous carcinoma	IIIc/3	66
529	papillary serous carcinoma	IIIc	67
533	papillary serous carcinoma	III/1	43
537	papillary serous carcinoma	IIIa/2-3	64
542	papillary serous carcinoma	IV/3	61
551	papillary serous carcinoma	IIIc/IV/3	59
559	papillary serous carcinoma	IV/3	49
588	papillary serous carcinoma	IIIc/2-3	71
589	papillary serous carcinoma	IIIc/IV/3	46
606	papillary serous carcinoma	IIIa/3	54
617	papillary serous carcinoma	IIIc/2-3	64
620	papillary serous carcinoma	III/IV/3	62
632	papillary serous carcinoma	IIIb/3	65
643	papillary serous carcinoma	IIIb/2	59
644	papillary serous carcinoma	IIIb/1-2	46
647	papillary serous carcinoma	IIIb-c/3	68
651	papillary serous carcinoma	IIIb-c/3	46
655	papillary serous carcinoma	III/IV/3	75
659	papillary serous carcinoma	IIIc/IV/3	78
678	papillary serous carcinoma	IV/3	59
688	papillary serous carcinoma	IIIc/3	59
694	papillary serous carcinoma	IIIb/3	70
704	papillary serous carcinoma	III/IV/3	75
717	papillary serous carcinoma	IIIb/3	64
721	papillary serous carcinoma	IIIc/1	58
756	papillary serous carcinoma	IIIc/2	59

Continued on the following page...

Table 1: Patients analyzed in this study

Patient ID	Ovarian Histopathology	Stage/Grade	Age at Surgery
782	papillary serous carcinoma	IIIc/3	59
787	papillary serous carcinoma	IIIc/3	72
821	papillary serous carcinoma	IIIc/1-2	58
831	papillary serous carcinoma	IIIc/	69
864	papillary serous carcinoma	IIIc/3	60
876	papillary serous carcinoma	IIa/1	63
5010	papillary serous carcinoma	IIIc/1	59
440	within normal limits	N/A	50
504	within normal limits	N/A	48
523	serous cystadenoma	N/A	32
534	within normal limits	N/A	72
540	within normal limits	N/A	59
541	within normal limits	N/A	41
544	within normal limits	N/A	49
552	within normal limits	N/A	41
612	within normal limits	N/A	48
614	within normal limits	N/A	44
615	within normal limits	N/A	42
623	simple cyst	N/A	54
627	within normal limits	N/A	59
636	within normal limits	N/A	71
650	cystic corpus luteum	N/A	47
677	within normal limits	N/A	68
691	within normal limits	N/A	70
693	simple cyst	N/A	60
697	within normal limits	N/A	51
698	functional cyst	N/A	49
703	within normal limits	N/A	42
719	fibrosis of tubal villi	N/A	55
733	within normal limits	N/A	37
736	within normal limits	N/A	45
737	within normal limits	N/A	41
740	functional cyst	N/A	37
749	simple cyst/cystic follicles	N/A	56
750	serous cystadenoma	N/A	41
751	within normal limits	N/A	60
752	within normal limits	N/A	74

Continued on the following page...

Table 1: Patients analyzed in this study

Patient ID	Ovarian Histopathology	Stage/Grade	Age at Surgery
755	within normal limits	N/A	75
757	within normal limits	N/A	84
759	within normal limits	N/A	52
763	hemorrhagic cyst	N/A	45
765	within normal limits	N/A	84
766	within normal limits	N/A	36
783	within normal limits	N/A	52
790	within normal limits	N/A	39
796	within normal limits	N/A	44
808	within normal limits	N/A	35
828	simple cyst	N/A	59
829	simple cyst	N/A	33
838	within normal limits	N/A	51
839	simple cyst	N/A	79
842	fibrosis of tubal villi	N/A	70
846	hemorrhagic corpus luteum	N/A	51
848	within normal limits	N/A	70
NHS1	healthy serum donor	N/A	36
NHS4	healthy serum donor	N/A	34
NHS10	healthy serum donor	N/A	37

2.2.2 Sample Preparation

Prior to analysis, 200 μL of each serum sample were thawed on ice and mixed with 1 mL of freshly-prepared, chilled (-18°C) and degassed 2:1 (v/v) acetone:isopropanol mixture. The mixture was vortexed and placed in a freezer at -18°C overnight to precipitate proteins followed by centrifugation at 13,000 g for 5 minutes. The supernatant was transferred to a new centrifuge tube, and the solvent was evaporated in a speed vacuum. The solid residue was re-dissolved in 25 μL of anhydrous pyridine (EMD Chemicals, Gibbstown, NJ), and shaken for one hour at room temperature for complete dissolution. Fifty μL of N-trimethylsilyl-N-methyltrifluoroacetamide (MSTFA, Alfa Aesar, Ward Hill, MA) containing 0.1% trimethylchlorosilane (TMCS, Alfa Aesar) were added to the sample in a N2-purged glove box. The mixture was then

incubated at 50°C in an inert N2 atmosphere for half an hour, resulting in TMS (tri-trimethylsilane)-derivatization of amide, amine, carboxyl and hydroxyl groups. The final derivatized mixture was subject to DART MS analysis.

2.2.3 DART-TOF MS

An in depth characterization of the analytical figures of merit of the DART MS approach used here has been recently reported [127], and therefore, the method is only briefly presented. Serum mass spectrometric analysis was performed using a DART ion source (IonSense Inc., Saugus, MA) coupled to a JEOL AccuTOF orthogonal time-of-flight (TOF) mass spectrometer (JEOL Inc., Japan). Prior to DART MS analysis, 0.5 μL of derivatized serum solution was pipette-deposited onto the glass end of the Dip-tip applicator (IonSense, Inc.) coupled to the sampling arm, a 1.2 min data acquisition run started, and the sample allowed to air dry for 0.65 min. The sampling arm was then rapidly switched so that the dried sample was exposed to the ionizing zone of the DART ion source. After 0.9 min in the acquisition run (0.25 min sampling time), the sample was removed, and a new Dip-tip placed on the sample holder, while the remaining 0.3 minutes of the run were completed. Each sample was run in triplicate.

The DART ion source was operated in positive ion mode with a helium gas flow rate of 3.0 L min-1 heated to 200°C. The glass tip-end was positioned 1.5 mm below the mass spectrometer inlet. The discharge needle voltage of the DART source was set to +3600 V, and the perforated, and grid electrode voltages set to +150 and +250 V, respectively. Accurate mass spectra were acquired in the range of m/z 60-1000 with a spectral recording interval of 1.0 s, and an RF ion guide peak voltage of 1200 V. The settings for the TOF mass spectrometer were as follows: ring lens: +8 V, orifice 1: +40 V, orifice 2: +6 V, orifice 1 temperature: 80°C, and detector voltage -2800 V. Mass drift compensation was performed after analysis of each sample using a 0.20 mM

polyethylene glycol 600 standard (PEG 600, Fluka Chemical Corp., Milwaukee, WI) in methanol. The measured resolving power of the TOF MS was 6000 at full width at half maximum, with observed mass accuracies in the range 2-20 ppm, depending on signal-to-noise ratios (S/N) of the particular peak investigated. A repeatability of 4.1-4.5% was obtained for the total ion signal using a manual sampling arm.

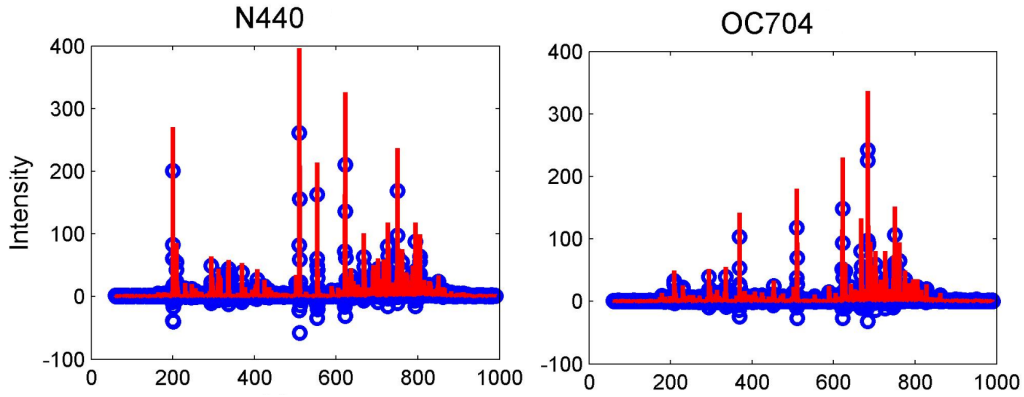
2.2.4 Data Preprocessing

All profile mass spectra were obtained by time averaging of the total ion chronogram between 0.73 and 0.76 minutes after each injection, that corresponds to the part of the time-varying signal that is conducive to the maximum number of analytes detected and identified with good sensitivity [127]. Following DART-TOF MS data collection and mass drift compensation, the background spectrum was subtracted and profile spectral data were exported in JEOL-DX format and converted to a comma-separated format prior to importing in MATLAB 7.6.0 (R2008a, MathWorks). The resulting data were normalized to a relative intensity scale and re-sampled to a total of 20,000 features between m/z 60 and 990 using the *msresample* function in the Matlab Bioinformatics Toolbox. The three replicate DART spectra were then averaged. The original dataset containing the DART-MS data can be downloaded [126].

2.3 Data Analysis

2.3.1 Functional Classification

In application domains such as chemometrics, it is well known that the shape of a spectrum is sometimes more important than its actual mean value. Therefore, it is beneficial to view the DART mass spectra as functions of m/z values (see Figure 4), and perform functional classification. The goal of functional classification [10] is to predict the label y of a functional data instance X given training data $S = \{X_i, y_i\}_{i=1}^{M}$, where the input functional data instance X_i is a random variable that takes values in an infinite dimensional Hilbert space H, the space of functions.



(a) Normal Patient (b) Cancer Patient

Figure 4: Mass Spectrum as Function

In functional data analysis, each observation X_{ij} of the data sample X_i arises from a smooth curve $x_i(t_j)$, t would be defined over time, space, wavelength, molecular weights and other continuums. Smoothness means that one or more of the curve's derivatives can be estimated. The observed data should provide enough information to estimate the function underlies the curve and its properties. Often, n discretization points have been chosen in $t_1, ..., t_N \in R$, and each functional data instance X_i is described by a vector . Sometimes, the functional data instances are badly sampled and the number and the location of discretization points are different between different instances. In practice, the functions that describe the input data instance will represented by constructing an approximation (such as B-spline interpolation) based on the observation values $((X_{i1}, \cdots, X_{iN} \in R^N)$, and then sampling uniformly to the reconstructed functional data $((x_i(t_1), \cdots, x_i(t_N)) \in R^N)$ [89].

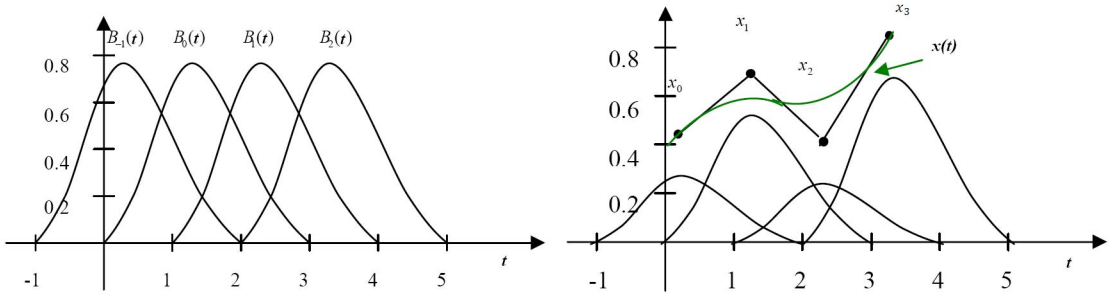
2.3.1.1 Basis Expansion

A natural way to model the smooth shape of each curve observed over t (time, molecular weight, etc.) is to choose a finite n-dimensional basis $\phi_k(t), k = 1, \cdots, n$.

$$x_i(t) = \sum_k c_{ik}\phi_k(t) = \Phi(t)^T c_i \tag{3}$$

where $\Phi(t) = [\phi_1(t), \cdots, \phi_p(t)]^T$ and c_i represents the basis coefficients for the ith

curve (corresponding to the ith instance). Given the functional building blocks, ϕ_k, and dimension, n, the basis coefficients c_i's can be estimated by apply standard linear least squares, separately to each curve. This fitting process from observation data X_i to function curve $x_i(t)$ is called as *basis expansion* (see Figure 5, black dots represent the observation values, the green curve represents the underlying function approximated through basis expansion).



(a) Example of Basis Functions (b) Example of Basis Expansion

Figure 5: Illustration of Basis Expansion

Common examples for basis functions include the Fourier, spline, and wavelet bases. If the functional data are known to be non-periodic, spline bases generally yield good results in practice [91]. We use B-Spline basis function in our experiments. The spline function is a polynomial of fixed degree or order over any subintervals of the observation interval, but the structure of the polynomial changes as one passes into the next sub-interval by matching a certain number of derivatives between adjacent sub-intervals. This class of basis functions are determined by:

- Number of sub-intervals (knots)
- Number of derivatives that are required to match at knot locations
- Degree or order of the polynomial segments (note that the order of a polynomial is its degree plus 1)

The number of basis functions in most spline setups is equal to the order of the polynomial segments plus the number of knots. We use equal-sized segment with degree 6 polynomials and second order derivative in our study, and curve flexibility is achieved by controlling the number of knots (i.e., sub intervals).

2.3.1.2 Functional Support Vector Machine

It has been suggested recently that the classification performance on functional data can be improved by designing SVMs specifically for functional classification [91, 111]. With the introduction of functional transformations and function kernels, functional SVM method can be described as follows:

1. Apply functional transformation, projection P_{V^n}, on each instance X_i as $P_{V^N}(X_i) = x_i = (x_{i1}, \cdots, x_{in})$ with X_i approximated by $\sum_{k=1}^{N} x_{ik}\Psi_k$, where x_{ik} is a complete orthonormal basis of the functional space H.

2. Build a standard SVM on the basis coefficients $x_i \in R^n$ for all $i = 1, \cdots, M$.

3. Apply projection P_{V^n} on the testing data point X, $P_{V^n}(X) = c$

4. Predict the label of the testing data point with the learned classifier with the functional coefficients $c \in R^k$ as input

This optimization procedure is equivalent to SVM optimization with a functional kernel, $K_n(x_i, x_j)$

$$K_n(x_i, x_j) = K(P_{V^n}(X_i), P_{V^n}(X_j)) \tag{4}$$

where P_{V^n} denotes the projection onto the n-dimensional subspace $V^n \in H$ spanned by $\{\Psi_k\}_{k=1,\cdots,n}$, and K denotes the standard linear SVM kernel.

2.3.2 Feature Filtering

The analysis of variance (ANOVA) is one of the most commonly used filter-based feature selection methods in bioinformatics. It helps to identify the features that highlight differences between groups [60, 61]. Let the dataset S contain c classes (groups), n data instances, and n_i instances from each class c_i; $X_{ij}(i = 1, \cdots, c; j = 1, \cdots, n_i)$ be a random sample of size n_i from a population with mean μ_i. ANOVA is used to investigate the null hypothesis $H0 : u_1 = u_2 = \cdots = u_c$ through F-test

$$f = \frac{SSB/(c-1)}{SSW/(n-c)} \quad , \tag{5}$$

where $SSB = \sum_{i=1}^{c}(\bar{x}_i - \bar{x})^2$ is the inter-class sum of square, $SSW = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$ is the total intra-class sum of squares, $\bar{x}_i$ and $\bar{x}$ are estimates of class and overall sample means respectively; x_{ij} is an observation (sample) from class c_i.

If the null hypothesis is rejected ($[f > F_{c-1,n-c}(\alpha)]$) , the upper αth percentile of the F-distribution with $c-1$ and $n-c$ degree of freedom), this implies that the groups of data samples differ significantly.

2.3.3 Evaluation Framework

The prediction performance of the DART-TOF MS dataset was evaluated through leave-one-out-cross-validation (LOOCV). Leave-one-out-cross-validation is generally considered as a more rigorous evaluation approach due to its maximal usage of the data for training [15, 14]. In this approach, at each 93-1 ($n = 94$ for our dataset) split cross validation, the chosen feature selection method was applied only to the training data, and then the prediction performance of the selected feature subset on the test data was measured. In our previous study on metabolite ovarian cancer biomarkers using LC-TOF (liquid chromagraphic - time of flight) MS technique, we show that this evaluation framework (illustrated in Figure 6) can avoid the selection bias in the prediction assessment [44].

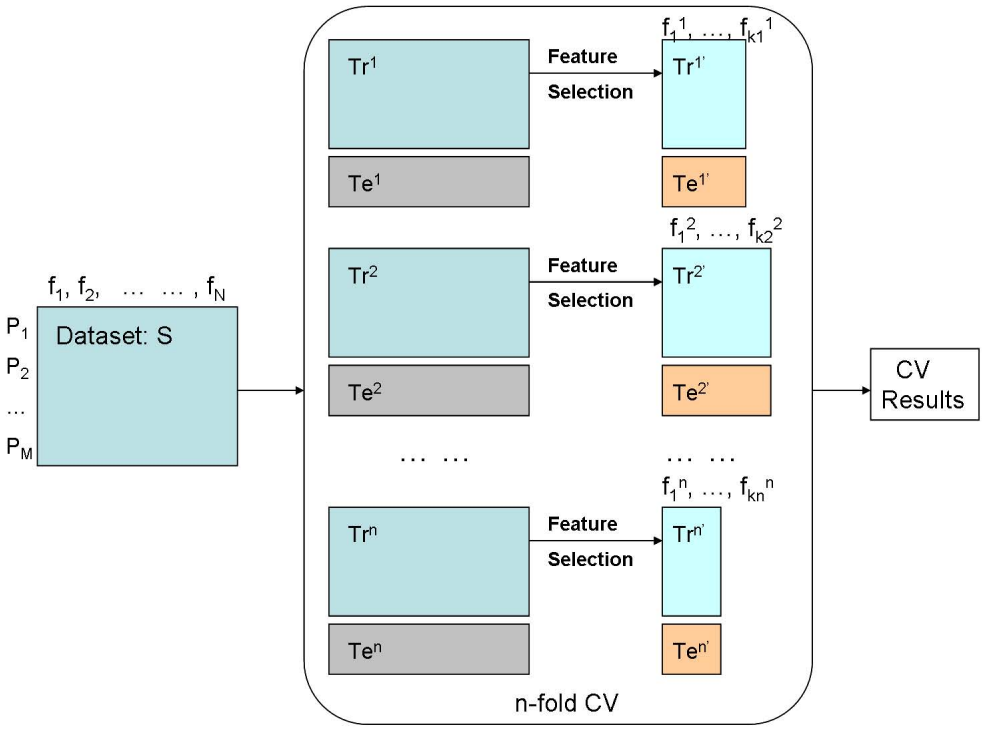


Figure 6: Evaluation Framework

SVM analysis of the DART mass spectra was performed using libSVM package [24]. fSVM analysis was performed using the FDA (functional data analysis) package [88] and libSVM. Partial least squares discriminant analysis (PLSDA) [8, 115] was performed using the PLS Toolbox (Eigenvector Research). We implemented the ANOVA feature selection method in Matlab.

2.3.4 Results Discussion

As described in Section 2, this ovarian cancer data set consists of DART MS metabolic profiles of 44 women diagnosed with serous papillary ovarian cancer (Stages I-IV) and 50 healthy women or women with benign conditions (e.g., serous, simple or follicular cysts). We analyze the data set with the customized Support Vector Machine (SVM) algorithm for classification of these metabolic profiles. The classification procedure can be briefly described as follows:

1. The data are collapsed along the desorption time dimension by using the average value within the time range of interest for all mass spectral m/z values ("features");

2. The resulting feature vector is smoothed using B-splines [89] to create the functional representation;

3. The vector of spline coefficients is then utilized by the support vector machine [109].

Table 2: Prediction Performance (Accuracy %) on the DART MS Data Set

Feature Selection	Features#	Classifier	SENS(%)	SPEC(%)	ACC(%)
One-way ANOVA ($\alpha = 0.05$)	4390	fSVM	**100**	**98**	**98.9**
		SVM	97.7	94	95.7
		PLSDA	97.7	**98**	97.9
One-way ANOVA ($\alpha = 0.01$)	2084	fSVM	**97.7**	**100**	**98.9**
		SVM	**97.7**	98	97.9
		PLSDA	93.2	92	92.6
every-7 sampling	2858	fSVM	**100**	**98**	**98.9**
		SVM	95.5	92	93.6
		PLSDA	93.2	90	91.5

For comparison purposes, we also evaluated the prediction performance of the data set using conventional techniques such as PLSDA (partial least square discriminative analysis) [8, 115] and SVM methods. In order to deal with the very large number of features (20,000 m/z values per serum sample run), the data were subjected to feature filtering before prediction performance evaluation. We compared the efficacy of the classifiers through leave-one-out cross-validation (LOOCV) evaluation. During LOOCV, each training set consisted of all patient samples except for one "left out" sample that is tested. In this way, each one of the patient samples is sequentially treated as an unknown, classified by the model as "cancer" or "control" in a blind fashion, and the accuracy of each classification evaluated. While validating models

by LOOCV, feature selection was performed independently on 94 different 93-1 split validations (see Figure 6).

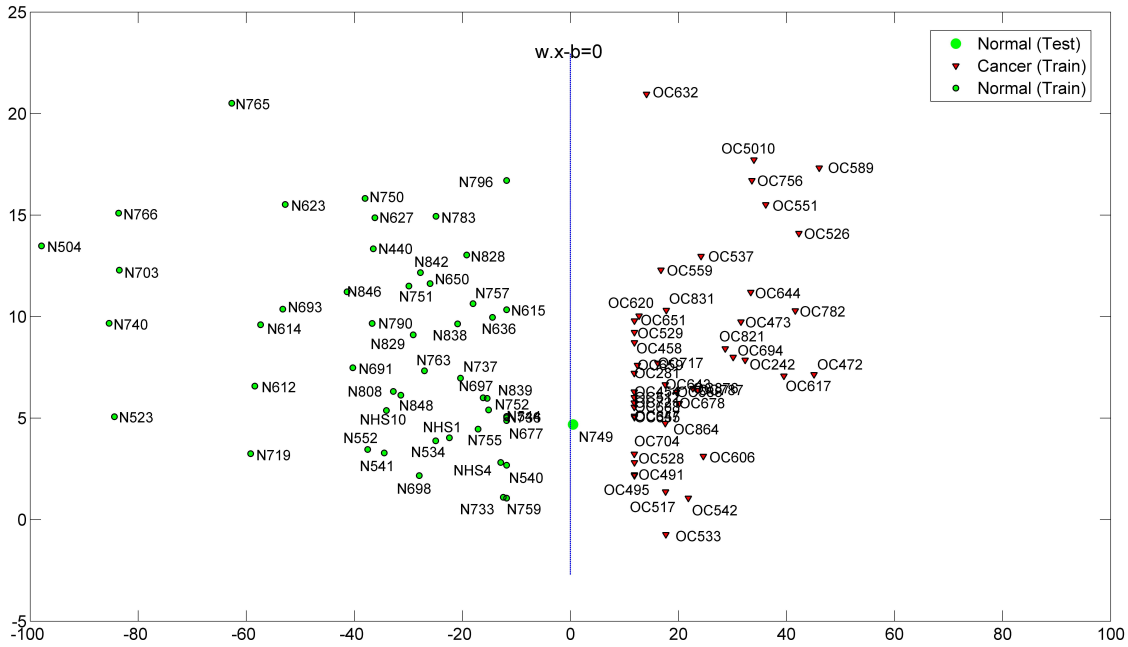


Figure 7: Visualization of a Functional SVM Classifier

As we can see from Table 2, all classification and feature selection methods showed high accuracy (over 90%) owing to the inherent discriminative power of the data. For example, combining with every-7-sampling feature filtering, only one 93-1 split validation of fSVM method resulted in a misclassification giving an overall accuracy of 99% (100% sensitivity and 98% specificity). The hyperplane of this 93-1 validation is visualized in Figure 7. The x axis is the optimal weight vector of the fSVM model. Red triangles with black edges correspond to ovarian cancer patients in the training set, green circles with black edges to controls in the training set, larger red triangles without borders are cancer patients in the test set, and the green circles without borders are the control samples in the test set. As shown in the figure, the testing data point in this validation, corresponding to control patient N749, which was misclassified but very close to the separating hyperplane. According to the patient information we collect, this patient has a very strong family history of ovarian cancer,

this probably can explain why this particular patient was misclassified by our method. Overall, the comparison study showed that fSVM method consistently outperforms the conventional methods under different combination of feature selection methods. The prediction performance of the conventional methods are more variational when combined with different feature selection methods.

2.4 Bilogical Validation and Discussion

2.4.1 Metabolite Identification

Features in the fSVM model utilizing every-7 sub-sampling (1:7:20,000) were assigned elemental formulae and tentatively matched to metabolites by finding the closest mass spectral peak matching the model features in the $[103, 714]$ m/z range. This m/z range was chosen because it is fully covered by the TOF calibration function thus providing the most reliable accurate masses. No attempt was made to match fSVM model features outside this range. Accurate masses were searched against a custom built database containing 2924 entries corresponding to elemental formulae of endogenous human metabolites in the HMDB database [53]. Each entry was manually expanded to take into account the mono, di and/or tri-trimethylsilane (TMS) derivatives. Entries for families of compounds not reacting with the MSTFA/TMCS reagent mixture were not expanded. Matching of database records to experimental DART MS data was performed using the SearchFromList application part of the Mass Spec Tools suite of programs (ChemSW, Fairfield, CA) using a tolerance of 10 mmu to obtain candidate elemental formulae. If no matches were found, the next closest match within 20 mmu was selected.

2.4.2 Pathway Enrichment Analysis and Metabolic Network Building

The MetaCore (GeneGO, St. Joseph, MI) software suite was used for metabolic network analysis. One hundred fifty-three estimated elemental formulae [126] obtained by DART MS accurate mass measurements of differentiating spectral features were

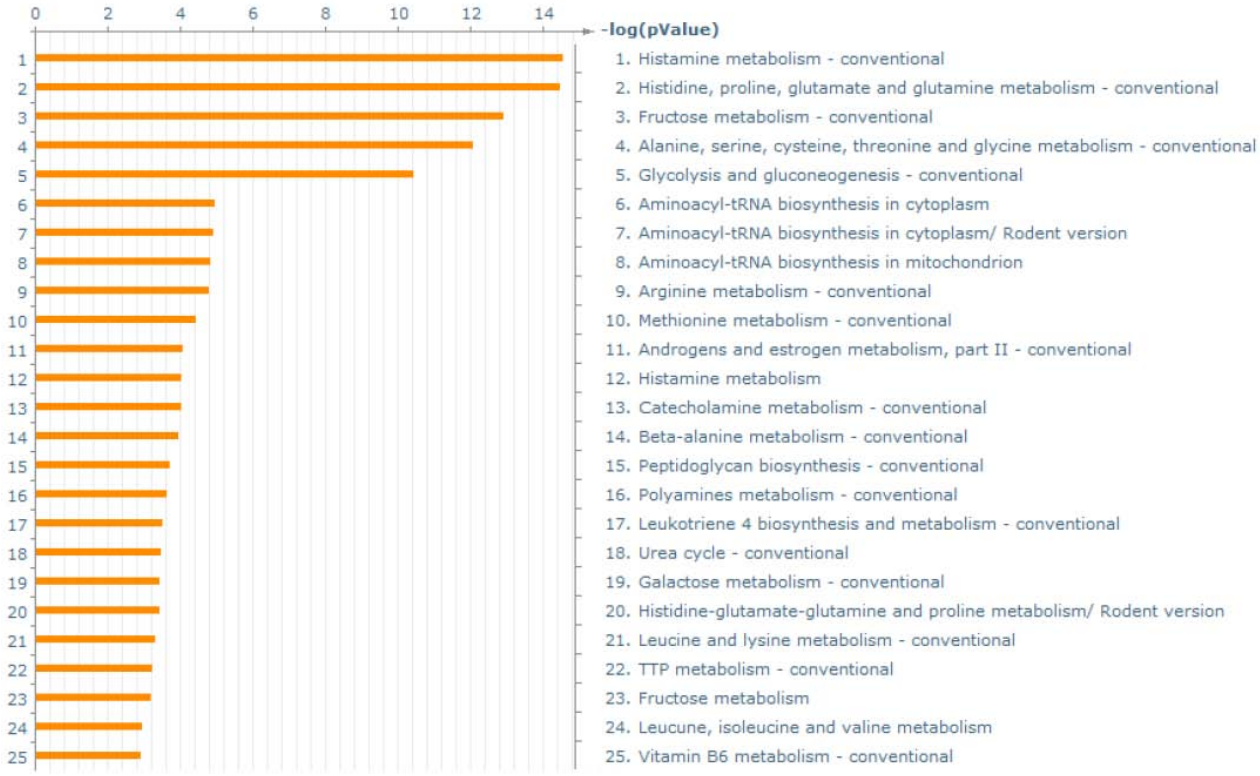


Figure 8: Canonical Metabolic Pathways Relevant to the Identified Metabolites

assigned to 385 network objects by the metabolic network analysis software, of which 299 represented unique endogenous metabolites or xenobiotic compounds [126].

Metabolic compounds assigned to these elemental formulae by MetaCore were mapped onto GeneGO canonical metabolic pathways that were ranked according to their relevance to the input set using p-values calculated based on hypergeometric distribution. These differentiating compounds can be mapped onto 25 pathways with p-values < 0.01. These 25 pathways are illustrated in Figure 8, and were ranked according to their p-values (hypergeometric distribution). This suggests differences between cancer and non-cancer groups in amine, amino acid, eicosanoid, and TTP (thymidine triphosphate) metabolism. Suggested differences in metabolism of carbohydrates and metabolism of androgens/estrogens have lower confidence since the relevance of corresponding pathways was determined from ambiguously identified metabolites (e.g., several different hexoses corresponding to elemental formula C6H12O6) and were not

further examined.

2.4.3 Potential Biological Significance of Metabolic Changes in Ovarian Cancer

A considerable proportion of the differentiating metabolites identified during the development of our assay represent components of the histamine pathway [126]. Serum histamine levels also have been reported to be altered in breast cancer [97]. Histamine is known to serve as a receptor-dependent growth factor in some colon, gastric, breast cancer and melanoma cell lines and to inhibit lymphocyte responsiveness via proliferation and activation of T lymphocyte suppressor cells [76]. In addition, the relationship of histamine with the metabolism of nitric oxide, polyamines and angiogenesis is an emerging area of interest in cancer biology [75]. The over-representation of members of the histamine pathway in our metabolic panel suggests that these species also may be of functional importance in ovarian cancer.

Other pathways over-represented in our dataset suggest that changes in the metabolism of several amino acids (e.g., glycine) involved in the de novo synthesis of purine nucleotides also are altered in ovarian cancer. Glycine, serine and sarcosine were all tentatively identified as differentiating metabolites in our study and these metabolites are components of over-represented canonical pathways of alanine, serine, cysteine, threonine, and glycine metabolism [126]. Several amino acids from these pathways previously have been identified in an earlier MS-based metabolic profile of ovarian cancer tissues [32]. Sarcosine, the N-methyl derivative of glycine, is elevated in invasive prostate cancer cell lines and in the tumors and urine of metastatic prostate cancer patients [101]. Also consistent with our findings, levels of these amino acids have all been previously reported to be elevated in colorectal [69], lung and breast [22] cancer patient sera.

A number of other tentatively identified metabolites (e.g., dopamine, tyramine, 5-hydroxykynurenamine and 1,2-dehydrosalsolinol) that are differentially expressed in

the sera of ovarian cancer relative to control patients are all products of decarboxylation of their precursor amino acids catalyzed by aromatic L-amino acid decarboxylase (DDC). This enzyme and its metabolic products previously have been shown to be elevated in neuroendocrine neoplastic tissues (carcinoid, small cell lung cancer; [22]). We recently have reported that DDC is overexpressed in ovarian cancer [12]. Networks built from our metabolic dataset using dopamine, tyramine, 5-hydroxykynurenamine and 1,2-dehydrosalsolinol and their precursors [126] are consistent with the finding that DDC (and its metabolic products) is (are) differentially expressed in ovarian cancer.

2.4.4 The Utility of Metabolic Profiling as a Diagnostic Test for Ovarian Cancer

Previous efforts to discover more accurate biomarkers of ovarian cancer using mass spectrometry have generally focused on large biopolymers, such as proteins [86]. However, finding and validating biomarkers of this kind has been plagued by the fact that the serum proteome is extremely complex, comprising $\sim 2 \times 106$ protein species with a dynamic range spanning 10 orders of magnitude [4]. This inherent complexity combined with current limitations in the proteomic analytical toolbox can result in the convolution of biomarker variability with non-biological sources of variance. Comprised of $\sim$2,500 molecules with molecular weights < 1000 Da, the known components of the serum metabolome can be readily distinguished from the serum proteome and more thoroughly interrogated [82]. As biological studies using more sensitive analytical tools with higher peak capacity improve our understanding of the serum metabolome, the number of detected and identified metabolites is expected to progressively increase, enriching the biological significance of discriminating spectral features useful in diagnostics.

MS analysis of serum samples typically employs chromatographic separation. This

step is usually time consuming and can result in increased costs and memory effects, which we believe was one of the confounding factors in our previous LC-MS study [44]. Our DART method circumvents chromatographic separation, making use of direct ionization without a matrix in a non-contact fashion. This decreases cross-contamination between experiments, enabling a better detection of differences between disease and control groups. Moreover, DART is able to ionize a broad range of metabolites with varying polarities [26], allowing simultaneous interrogation of multiple chemical species at minimal cost.

By combining the DART-TOF MS with a customized fSVM classification algorithm, we were able to distinguish sera from cancer patients and controls with 99% accuracy using the LOOCV test (100% sensitivity and 98% specificity). In this study, the use of high resolution TOF MS was necessary for metabolite identification purposes, but the spectral data were later down-sampled for machine learning purposes, suggesting that approaches similar to the one presented here, but based on low resolution MS data acquisition, may also be conducive to high discriminatory power.

There is general consensus among the ovarian cancer community that to be of clinical significance, a diagnostic test for ovarian cancer must have a minimum positive predictive value (PPV) of 10% [96]. Because the prevalence of ovarian cancer in the general population is low (0.04%), the accuracy of any potential screening test to be used in the general population must be extremely high (100%) [57]. While our results indicate that our approach has great potential as a diagnostic tool of clinical significance, more extensive testing will be required to define its use in screening applications. Other, more immediate clinical applications of our assay may be in those sub-populations of women where the prevalence of ovarian cancer is known to be relatively high. For example, the estimated incidence of ovarian cancer in women aged 20 and over with 2 first-degree relatives with ovarian cancer is 0.266% [18]. Using incidence to approximate prevalence [103], we estimate a clinically significant

12% PPV for our assay in this sub-population assuming the LOOCV values of 100% sensitivity, 98% specificity. Women 20 years of age and over who test positive for BRCA1 or BRCA2 are reported to have an incidence of ovarian cancer as high as 0.683% [108]. For this group of women, our assay would have an estimated PPV of 26% – well above the minimum value (10%) for a test to be considered of clinical significance.

The results presented here demonstrate the potential application of our method as an ovarian cancer diagnostic of significant clinical value. In addition, if future studies establish that metabolic profiles of different cancers and other diseases are sufficiently distinct, our method may have the added advantage that it could be used to rapidly and inexpensively test for multiple diseases from a small serum sample.

CHAPTER III

BIOMARKER DISCOVERY FROM HIGH THROUGHPUT CANCER DATA

3.1 Introduction

Biomarkers, a consensus definition of which is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [6, 112], are crucial for successful disease diagnosis and treatment development.

High throughput technologies such as microarray, mass spectrometry, etc., provide simultaneous evaluation of thousands of potential biomarkers that distinguish between different patient groups, e.g. normal versus cancer patients. For example, Affymetrix U95Av2 microarray chip measures expression value of 12,558 genes per patient; intensity values of 20,000 or more different m/z values can be extracted from a DART-TOF mass spectrum of a patient. Because of the high experimental cost, high throughput cancer data sets are usually low sample size, involving only a few hundreds patients. Therefore, the high throughput cancer data are typically low-sample-size and high-dimensional.

Biomarker discovery, also referred as gene selection, metabolite panel selection, allows to make the diagnostic process cheaper and targeted, and to narrow down the number of biomarkers to better understand their biological significance. The task is typically reduced into classification-driven feature selection and subsequent biological validation. An example of the biological validation process on the cancer biomarker candidates is described in Section 2.4. Here, we focus on developing more aggressive classification-driven feature selection methods that are suitable for these low-sample

size high-dimensional cancer data sets.

A key difficulty for biomarker discovery such high-throughput cancer data is very noisy, which can be caused by the intrinsic complexity of the biological processes related to cancer, as well as experimental and technical imperfections. Another difficulty arises from the low sample size and high dimensionality of the data. There is a high risk of overfitting of standard feature selection method due to the small sample size. Thus, there is an algorithmic need for developing feature selection for these high throughput cancer data sets.

In this work, we address feature selection in the context of linear support vector machine (SVM) learning [109]. For a two-class classification problem on a data set $S = \{x_i, y_i\}_{i=1}^n$, where $x_i \in R^d$ is the feature vector of the ith data point, and $y_i \in \{-1, +1\}$ is the corresponding label, we can formulate SVM learning into the following optimization problem:

$$\min_{\mathrm{w},\gamma} \mathrm{C} \sum_{\mathrm{i}} \mathrm{L}\{\mathrm{f}(\mathrm{x_i}), \mathrm{y_i}\} + \frac{1}{2} \sum_{\mathrm{j}} \mathrm{w_j^2} \quad , \tag{6}$$

where $L\{f(x), y\} = [1 - yf(x)]_+$ is the hinge loss, $f(x) = wx - \gamma$ is the decision function, w is the weight vector, γ is the bias term. Tuning parameter $C > 0$ controls the trade-off between the goodness of data fit, as measured by the hinge loss, and the complexity of model f, as measured by the L_2-norm of the weights.

Existing approaches to feature selection for SVMs mainly fall into three categories: filter-based methods, wrapper-based methods, and embedded methods. Filter-based methods adopt feature ranking strategies disjoint from SVM training, such as t-statistics, signal-to-noise ratio, etc. This type of feature selection methods usually compute ranking score of each individual feature according to certain ranking criterion and then select out the k best features (sorted by ranking scores). Although they might be preferable because of computational efficiency and statistical robustness, individual feature ranking is far from optimal in many cases [46].

In real-world data sets, it's very common that a feature that is useless by itself can provide a significant prediction performance improvement when combined with some others features. Therefore subset feature selection is desirable in spite of its NP-hardness [3]. Wrapper-based methods assess the relative importance of feature subsets based on their SVM hyperplane parameters or SVM performance on the training data. Representative methods of this approach include recursive feature elimination [48], and Weston's R2W2 method [113], etc. Current wrapper methodologies usually adopt greedy search strategies, such as forward selection and backward elimination, to avoid exhaustive brute force feature subset search. However, these methods could still be computational expensive.

Embedded methods augment the SVM formulation, and seek to learn the SVM classifier as well as the optimal feature subset simultaneously in one optimization. Significant examples of this approach include L_1-norm SVM (L_1-SVM) [13], Feature Selection concaVe (FSV) [13, 67], etc. Embedded methods incorporate subset feature selection as part of the SVM learning process, thus they might be more computational efficient than wrapper-based feature selection methods.

In this work, we focus on deriving feature selecting support vector machine from L_0-norm SVM formulation. Unlike L_1-norm SVM, which performs feature selection as a by-product because of the resulting sparse solution, L_0-norm SVM directly minimizes on both hinge loss and cardinality of its weight vector. In the context of regression, where feature selection has been most throughly studied, it has been pointed out that though L_1 penalization yields sparse solutions, the estimates can be biased since larger penalties are imposed on larger weight coefficients [34]. A recent comparison study of least absolute shrinkage and selection operator (LASSO) regression [106] and forward stepwise regression, which is a greedy surrogate of L_0-regularization, further indicated that L_1-regularization never outperforms L_0-regularization by more than a constant factor, and in some cases, using an L_1-norm penalty is much worse than an

L_0-norm penalty [72]. This comparison analysis also pointed out that "an approximation solution to the right problem can be better than the exact solution to the wrong problem" [72]. Our study follows this guideline.

The rest of this chapter is organized as follows: In the next section, we briefly summarize several widely-used embedded feature selection methods. In section 3, we describe the mixed-integer SVM problem and its convex relaxation formulated as mixed-integer quadratic problems, and then present the comparison study on six real-world data sets. To better address the scalability issue on the low sample size high dimensional cancer data sets, we further propose the fractional-norm SVM problem and solve it through difference of convex programming technique. The optimization method and the results analysis on synthetic data sets and seven real-world data sets are described in Section 4.

3.2 Related Work

3.2.1 L_1-norm SVM

Bradley and Mangasarian (1998) [13] proposed the L_1-norm SVM (L_1-SVM) method, which solves the following optimization problem.

$$\begin{aligned} \min \quad & \| w \|_1 + c \| \xi \|_1 \\ \text{s.t.} \quad & Y(Xw - \gamma e_m) + \xi \geq e_m \\ & \xi \geq 0 \end{aligned} \tag{7}$$

L_1-norm SVM method performs feature selection as a by-product of the resulting sparse solution. Defining $w = p - q$, with $p, q \geq 0$, Equation (7) is then equivalent to the linear programming problem below,

$$\begin{array}{lrcl} \min & e_n^T(p+q) + ce_m^T\xi & & \\ & YXp - YXq - \gamma Ye_m + \xi & \geq & e_m \\ \text{s.t.} & p, q & \geq & 0 \\ & \xi & \geq & 0 \end{array} . \tag{8}$$

Several efficient algorithms are proposed for for L_1-SVM optimization, such as Fung and Mangasarian (2004) [37]; Zhu et al. (2004) [128]; Mangasarian (2007) [74], etc. Because of its computational efficiency and the empirically sparse solutions, L_1-norm SVM method and its variants have been applied to various problems in computation biology [44] and many other domains.

3.2.2 L_0-norm SVM and Its Approximation Methods

Unlike L_1-norm SVM method, L_0-norm SVM directly minimizes on both hinge loss, $\sum_i [1 - y_i(wx_i - \gamma)]_+$, and cardinality of its weight vector. However, optimizing the L_0-norm SVM is a NP-hard problem [3]. Previous work in this direction mainly includes adopting smoothed approximations of the L_0-norm, using adaptive scaling parameters to control the sparsity. We select representative methods from each category for our comparison study.

Weston et al. (2001) [113] introduced the idea of scaling variable, a feature is removed if the corresponding scaling variable becomes zero during the optimization. This method learns the optimal scaling variables and SVM classifier through minimizing a generalization error bound, R^2W^2,

$$\begin{array}{rcll} R^2(\beta,\pi) & = & \min & \beta^T X\Pi X^T\beta - \sum_{i=1}^n \beta_i x_i^T \Pi x_i \\ & & \text{s.t.} & e_n^T\beta = 1, \beta \geq 0 \\ W^2(\alpha,\pi) & = & \min & \frac{1}{2}\alpha^T YX\Pi X^T Y\alpha - e_n^T\alpha \\ & & \text{s.t.} & \alpha^T y = 0, \alpha \geq 0 \end{array} , \tag{9}$$

where $R^2(\beta,\pi)$ minimizes the radius of the smallest sphere, centered at the origin,

that contains all the data points; $W^2(\alpha, \pi)$ maximizes the margin distance of the learned classifier, matrix $\Pi = \text{diag}\{\pi\}$ and $\pi \in \{0,1\}^d$ denote the scaling variables. The above problem is relaxed to $\pi \in R_+^d$ in the approximation algorithm. At each iteration t, the algorithm first optimizes $R^2(\beta, \pi^{(t-1)})$ to get α^t, and $W^2(\alpha, \pi^{(t-1)})$ to obtain β^t. Second, it updates π with gradient of $R^2(\beta^{(t)}, \pi)W^2(\alpha^{(t)}, \pi)$. Third, it sets the smallest nonzero π_k^t to zero, i.e. discards the corresponding feature. The above procedure repeats until only d features left. We denote this method as R2W2.

Bradley and Mangasarian (1998) [13] approximated the L_0-norm penalty, $\sum_{j=1}^d (w_j)^0$ with smooth function $\sum_{j=1}^d (1 - \exp^{-\alpha|\mathrm{w_j}|})$. The resulting approximation of L_0-SVM is the Feature Selection concaVe (FSV) problem,

$$\begin{aligned} \min J1(w, r, y, z, v) \quad & = (1-\lambda)(\tfrac{e^T y}{n_+} + \tfrac{e^T z}{n_-}) + \lambda e^T (e - \exp^{-\alpha|\mathrm{w}|}) \\ & -Aw + e\gamma + e \leq y \\ \text{s.t.} \quad & Bw - e\gamma + e \leq z \\ & y \geq 0, z \geq 0 \end{aligned} \quad , \tag{10}$$

where A represents the data matrix of the positive training samples, and B is the data matrix of the negative training samples.

Bradley and Mangasarian (1998) [13] solve Problem (10) using Successive Linearization Algorithm (SLA), which iteratively update solution through optimizing the linear programming problem below

$$\begin{aligned} \min \quad & (1-\lambda)(\tfrac{e^T y}{n_+} + \tfrac{e^T z}{n_-}) + \lambda\alpha(\exp^{-\alpha \mathrm{v^t}})^{\mathrm{T}}(\mathrm{v} - \mathrm{v^t}) \\ & -Aw + e\gamma + e \leq y \\ & Bw - e\gamma + e \leq z \\ \text{s.t.} \quad & -v \leq w \leq v \\ & y \geq 0, z \geq 0 \end{aligned} \quad . \tag{11}$$

where $(w^t, \gamma^t, y^t, z^t)$ is the solution at the tth iteration. We denote this method as FSV.

Le Thi et al. (2008) [67] solved the FSV problem with DC programming. At each iteration, they optimizes a linear programming problem based on DC decomposition, $J1(\cdot) = G1(\cdot) - H1(\cdot)$. We denote this method as DC-FSV.

$$\begin{array}{rcl} G1(w,\gamma,y,z) & = & (1-\lambda)(\frac{e^T y}{n_+} + \frac{e^T z}{n_-}) + \lambda \sum_{j=1}^{d} \alpha \mid w_j \mid \\ H1(w,\gamma,y,z) & = & \lambda \sum_{j=1}^{d} (\alpha \mid w_j \mid -1 + \exp^{-\alpha \mid w_j \mid}) \end{array} \quad . \tag{12}$$

Fan and Li (2001) [34] proposed the smoothly clipped absolute deviation (SCAD) penalty (Eqn. (13)) to approximate the L_0-norm penalty,

$$p_\lambda(|w|) = \begin{cases} \lambda |w| & \text{if } |w| \leq \lambda \\ -\frac{|w|^2 - 2a\lambda|w| - 1}{2(a-1)} & \text{if } \lambda < |w| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |w| > a\lambda \end{cases} \quad . \tag{13}$$

Zhang et al. (2005) [120] solved the SCAD-penalty regularized SVM problem (Eqn. (14)) using successive quadratic algorithm (SQA).

$$\min_{b,w} \frac{1}{n} \sum_{i=1}^{n} [1 - y_i(b + w \cdot h(x_i))]_+ + \sum_{j=1}^{d} p_\lambda(|w_j|) \tag{14}$$

This is an iterative algorithm: at each iteration, it optimizes an unconstrained quadratic programming problem derived from the quadratic approximation (Eqn. (15)) of the SCAD penalty,

$$\begin{array}{rcl} [1 - y_i(b + w \cdot x_i)]_+ & = & \frac{1 - y_i(b + w \cdot x_i)}{2} + \frac{|y_i - (b + w \cdot x_i)|}{2} \\ p_\lambda(|w_j|) & \approx & p_\lambda(|\hat{w}_j|) + \frac{p'_\lambda(|\hat{w}_j|)}{2|\hat{w}_j|}(w_j^2 - \hat{w}_j^2) \end{array} \quad , \tag{15}$$

where $\hat{w}_j$ is the optimal solution from the previous iteration. We denote this method as SCAD-SVM.

Liu et al. (2007) [73] proposed local quadratic approximation (LQA) algorithm to optimize the L_q-SVM problem (Eqn. (20)). At each iteration, the algorithm optimizes an unconstrained quadratic programming problem derived from the quadratic approximation on the L_q-norm penalty,

$$| w_j |^q \approx | \hat{w}_j |^q + \frac{(| \hat{w}_j |^q)'}{2 | \hat{w}_j |}(w_j^2 - \hat{w}_j^2) \quad , \tag{16}$$

where $\hat{w}_j$ is the solution from the previous iteration. We denote this method as LQA.

3.3 Mixed-Integer SVM

In contrast to previous work, which either used a smoothed penalty function that approximates the L_0-norm in the objective or used adaptive scale parameters, and then solved through convex optimization techniques, we present the mixed-integer support vector machine (MI-SVM) based on mix-integer relaxations on the L_0-SVM formulation, and then optimize the problem using mixed-integer nonlinear programming (MINLP) techniques. Empirical comparison of our MI-SVM method with the standard SVM, L1-SVM, FSV, and Weston's R2W2 feature selection methods, demonstrates either sparser solutions with roughly identical classification performance, or an increase in classification performance with similar or sparser representations.

3.3.1 Mixed-Integer Relaxations on L_0-norm SVM

We consider the following L_0-norm SVM formulation:

$$\begin{aligned} \min_{w,\gamma,\xi} \quad & || w ||_0 + c || \xi ||_1 \\ \text{s.t.} \quad & Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0 \end{aligned} \tag{17}$$

Inspired by Gunluk and Linderroth's work on perspective relaxation of indicator-induced MINLP problems [45], we relax Problem (17) by introducing indicator variable $z_j \in \{0, 1\}$, $z_j = 0 \Rightarrow w_j = 0$ and $z_j = 1 \Rightarrow w_j > 0$; and the perspective constraints $w_j^2 \leq z_j u_j$, where u_j is the squared upper bound of the weight element w_j. These define a convex hull of $w_j^2 = z_j u_j$, which is the equality we want to enforce. The proposed mixed-integer SVM can then be formulated as the following mixed-integer quadratically constrained quadratic program.

$$
\begin{aligned}
\min_{z,u,w,\gamma,\xi} \quad & a e_n^T z + \tfrac{1}{2} e_n^T u + c e_m^T \xi \\
\text{s.t.} \quad & Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0 \\
& w_j^2 - z_j u_j \leq 0 \\
& z \in \{0,1\}^n,\ u \geq 0,\ \ \xi \geq 0
\end{aligned}
\tag{18}
$$

where vector $z = [z_1, \cdots, z_n]^T$, $u = [u_1, \cdots, u_n]^T$, and constants $a, c > 0$ adjust the trade-off between the cardinality of the weight vector and the hinge loss.

The above equation tries to minimize the L_0-norm penalization $\sum_j z_j$, the L2-norm upper bound $\sum_j u_j$, and the hinge loss $\sum_i \xi_i$. The first type of constraints $Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0$ regulates the classification error for each training instance. The second type of constraints $w_j^2 \leq u_j z_j$ enforce that i) $w_j = 0$ when $z_j = 0$, and ii) $u_j = w_j^2$ at optimal.

However, solving this problem directly with the existing MINLP tools such as Bonmin [11] or MINLP [70] fails. The experiments of optimizing this problem over even small datasets resulted in either infeasible states or unsatisfying solutions with all indicator variable setting to zero. We believe that the failure of the nonlinear solvers is due to a failure of constraint qualification on the conic constraints $w_j^2 \leq z_j u_j$. For example, whenever $z_j = 0$ during the tree-search or in the solution of continous subproblems in Bonmin, the relaxation contains a constraint $w_j^2 \leq 0$, which violates Slater's constraint qualification [52]. While it is in principle straightforward to remedy this situation by preprocessing the constraint $w_j^2 \leq 0$ and replacing it by $w_j = 0$, current nonlinear solvers do not perform this operation. The errors that we observe from the nonlinear solvers are consistent with a failure of a constraint qualification.

To remedy this adverse situation, we thus relax the conic constraints $w_j^2 \leq u_j z_j$ into big-M constraints $\mid w_j \mid \leq M z_j$, where M is a fixed large number (M was set to 10^4 in our experiments). This results in a mixed-integer quadratic problem (19) .

$$
\begin{aligned}
\min_{z,w,\gamma,\xi} \quad & a e_n^T z + \tfrac{1}{2} w^T w + c e_m^T \xi \\
\text{s.t.} \quad & Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0 \\
& \mid w_j \mid - M z_j \leq 0 \\
& z \in \{0,1\}^n,\ \xi \geq 0
\end{aligned}
\tag{19}
$$

3.3.2 Results and Discussion

We compare the performance of the proposed MI-SVM method, with the standard SVM, L_1-SVM, FSV and Weston's R2W2 feature selection methods on six real-world biomedical data sets. We use the LibSVM package [24] for standard SVM optimization and the L1-SVM code from [37, 74]. We implemente the FSV, and R2W2 methods in MATLAB. Due to numerical reasons, for FSV method, the elements of the optimal weight vector that have small relative magnitude, i.e. $\frac{|w_j|}{max_k(|w_k|)} < 10^{-4}$, are set to zero. We use CPLEX [55] to solve the mixed-integer SVM problem. Since the CPLEX tool has the similar numerical issues, we can apply the same threshold rule as the FSV method. We denote this approach as MI-SVM^1. Furthermore, since we also obtain the optimal indicator variable assignment after solving Eqn. (19), we would apply standard SVM on the subset of the data that only contains the features with non-zero indicator variables, and then obtain the final weight vector. We denoted this approach as MI-SVM^2.

3.3.2.1 Data Sets

We select four widely-used real world biomedical data sets from the UCI repository [78] for our experiments. *Ionosphere* data set consists of 351 instances with 34 features. There are 225 radar returns termed "good" or showing some type of structure in the ionosphere, and 126 radar returns termed "bad"; their signals pass through the ionosphere. *Wisconsin prognostic breast cancer (wpbc)* data set consists

of 198 instances with 32 numerical features representing follow-up data of the patients. Two of its variants are used. The first data set (denoted as wpbc24) includes 28 patients who had a cancer recurrence in less than 24 months and 127 patients who didn't have a cancer recurrence in less than 24 months. The second variant (denoted as wpbc60) contains 41 patients with a cancer recurrence in less than 60 months, and 69 patients which cancer had not recurred in less than 60 months. *SPECTF heart* data set: the training dataset consists of 80 instances with 44 features (40 instances labeled with "1" and "0", respectively); the testing dataset consists of 187 instances with 172 instances labeled with "1" and 25 labeled with "0".

We also use sub datasets of our metabolomic mass spectrometry cancer data sets in our experiments. The *OvarianCancer* data set [44] consists of metabolomic profiles of 37 ovarian cancer patients and 35 benign controls. Each metabolomic profile contains intensity values of the same 592 features extracted from the LC/TOF mass spectra of the patient serum. 360 of the 592 features are in the pos-ion-mode and the remaining 232 features are in the neg-ion-mode. The *DART* data set [126] consists of metabolomic profiles of 44 women diagnosed with serous papillary ovarian cancer (stages I-IV) and 50 healthy women or women with benign conditions. The metabolomic profiles are obtained using DART mass spectrometry technique. Each metabolomic profile contains the intensity values of 20,000 features that are uniformly resampled within m/z range [60, 990] based on the normalized raw DART-TOF mass spectra. To reduce the curse of dimensionality, for each data set, we filter the features according to t-statistics and select out the top 50 features for our study.

3.3.2.2 Parameter Tuning

We estimate the generalization ability of each method via 10-fold cross validation (10-fold CV) on each data set, except for *SPECTF* data set as its training and testing split are given. Note that we need to tune the parameter c of the standard

SVM method and the RFE method, parameters δ, c of L1-SVM, parameter λ for FSV method, and parameters a, c of the MI-SVM methods for the performance evaluation. We employ the following parameter tuning procedure on each data set: for each parameter setting, we perform a 10-fold CV, and the score for this parameter setting is the averaged training accuracy over cross-validation. For *SPECTF* data set, we use the training accuracy as its score. Then we select the parameter setting with the best score (ties are broken by choosing the sparser solutions). The candidate parameter values used in the experiments were

- $c \in \{2^{-7}, \cdots, 2^{-1}, 1, 2^1, \cdots, 2^7\}$,
- $\delta \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\}$,
- $\lambda \in \{0.05, 0.1, 0.15, \cdots, 0.85, 0.9, 0.05\}$,
- $a \in \{2^{-3}, \cdots, 2^{-1}, 1, 2^1, \cdots, 2^3\}$.

3.3.2.3 Result Analysis

Table 3: Feature Selection Performance (Number of Selected Features)

	SVM	R2W2	FSV	L1-SVM	MI-SVM^1	MI-SVM^2
OvarianCancer-50	50	49	17	24	17	**13**
DART-50	50	13	**5**	8	10	10
Ionosphere	33	30	30	30	31	31
wpbc24	32	27	**5**	24	19	24
wpbc60	32	22	19	25	21	**16**
SPECTF	44	21	34	28	**12**	**12**
on-average	40	27	18	23	18	**18**

Table 14 summarizes the feature selection performance, measured by the number of features selected by each method. Table 4 describes the classification performance, measured by testing accuracy of each classifier. Overall, the experiment results show that MI-SVM methods are able to learn sparser representations with roughly identical or increased classification performance. And MI-SVM^2, the approach of applying

Table 4: Classification Performance (Accuracy %)

	SVM	R2W2	FSV	L1-SVM	MI-SVM^1	MI-SVM^2
OvarianCancer-50	66.3	64.8	67.7	63.2	63.8	**76.1**
DART-50	78.8	91.4	96.7	95.7	95.7	95.7
Ionosphere	88.4	**88.7**	88.1	**88.7**	**88.7**	**88.7**
wpbc24	78.8	78.2	70.8	81.5	81.3	**82.1**
wpbc60	66.2	**66.5**	61.9	58.2	60.9	63.6
SPECTF	72.2	73.8	73.8	58.8	**76.5**	**76.5**
on-average	75.1	77.2	76.5	74.3	77.8	**80.4**

standard SVM onto the feature selection results of Eqn. (19) (features with non-zeros indicator variables) had a higher prediction performance than MI-SVM^1, the approach of simply thresholding out the optimal weights of Eqn. (19) that having relative small magnitude.

MI-SVM^2 approach outperforms the other compared method with average testing accuracy of 80.4% and average selected feature size of 18 over the six data sets. MI-SVM^1 and R2W2 methods had the second best testing accuracy (77.8% on average) and MI-SVM^1 were able to find sparser representation than R2W2 method. While L1-SVM had the worst prediction performance (74.3% averaged testing accuracy) with average of selected feature size of 23. In data sets such as *wpbc24*, *SPECTF*, *Ovarian*, and *DART*, the testing accuracy increase significantly when using MI-SVM, which indicates that some of the features in these datasets may be irrelevant to the disease. In other data sets like *Ionosphere*, *wpbc60*, the prediction accuracy remains roughly the same while sparsity increases, which suggests that these data sets may contain redundant features. In both cases, our MI-SVM method consistently learns lower dimensional representations of the data sets with improved or comparable classification performance. This demonstrates the effectiveness of MINLP techniques which have not previous been widely used in machine learning.

3.4 Fractional-norm SVM

Despite the improvement on prediction and feature selection performance, MI-SVM optimization is very computationally expensive and the method is not applicable to high-dimensional (or even medium feature size) data sets. To better handle the scalability issue, we further propose the fractional-norm SVM problem, which achieve sparsity by augmenting the SVM objective function with L_q-norm penalty term, for $q \in (0, 1)$. In this section, we describe our optimization solution to the fractional-norm SVM problem with the Difference of Convex functions (DC) programming technique [85]. DC programming firstly decomposes a non-convex objective function into the difference of two convex functions, and then solves the resulting problem through a primal-dual approach. Under such a framework, we present an iterative algorithm scheme for the fractional-norm SVM learning problem, which at each iteration solves a reweighted L_1-SVM problem. Therefore, we can reuse the existing efficient optimization methods for L_1-SVM in our algorithm. We also give some theoretical convergence guarantees about the algorithm. The empirical study with several popular sparse SVM methods indicates that the proposed DC programming approach leads to better performance in both classification and feature selection, with good computational efficiency, especially on low sample size high dimensional data sets.

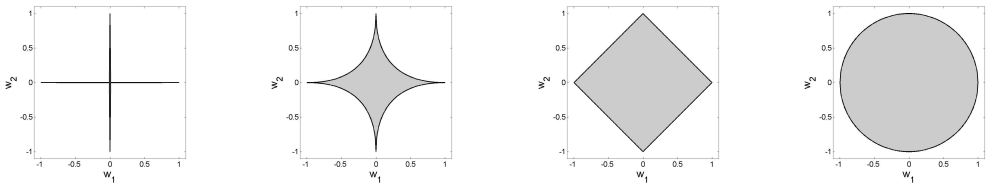
3.4.1 L_q-norm Regularization

L_q-norm penalty is defined as $R_q(w) = \sum_j \mid w_j \mid^q$, for $q > 0$. Previous analysis on the effect of L_q-norm regularization (also called as bridge penalty) in the context of regression [66, 34, 129, 38], where feature selection has been most thoroughly studied, indicated that:

- When $q \geq 1$, the larger q is, the more penalties are imposed on coefficients with $\mid w_j \mid > 1$, and the less penalties are imposed on coefficients with $\mid w_j \mid < 1$. If $q > 1$, the L_q regularization does not threshold coefficients. If $q = 1$, small

$| w_j |$'s are tends to shrink to zero (i.e. achieving feature selection).

- When $0 < q < 1$, the smaller q is, the more penalties are imposed on coefficients with $| w | < 1$, and the less penalties are imposed on coefficients with $| w | > 1$. The L_q regularization may achieve better sparsity than the L_1 regularization because larger penalty is imposed on small coefficients than the L_1 regularization.



(a) q=0.1 (b) q=0.5 (c) q=1 (d) q=2

Figure 9: Feasible Region of L_q Regularization $|w_1|^q + |w_2|^q \leq 1$

In the context of classification, especially SVM learning, the L_q-norm SVM problem (denoted as L_q-SVM) can be formulated as follows:

$$\min_{w,\gamma} C \sum_{i} L\{f(x_i), y_i\} + R_q(w) \quad . \tag{20}$$

where hinge loss $L(f(x_i), y_i) = [1 - y_i f(x_i)]_+$, decision function $f(x_i) = wx_i - \gamma$.

With $q = 2$, this is the standard SVM, which generally utilizes all the input features in the learned decision function. With $q = 1$, this is the L_1-SVM, which was shown to perform reasonably well in many situations [13]. With $q = 0$, this is the L_0-SVM, which directly enforces sparsity with the cardinality of w, however, L_0-SVM optimization has been shown to be NP-hard [3].

Here, we consider the fractional-norm SVM (i.e., L_q-SVMs with $q \in (0, 1)$), due to its promise as a middle ground between L_0 (the most aggressively-sparse, but intractable) and L_1 (less sparse but much more tractable). For small values of q, it can be seen as a quasi-smooth approximation of L_0. Though promising, L_q regularization

presents a much more difficult optimization problem than does L_2 or L_1, as the objective function is non-convex and not differentiable at zero. In this work, we develop practical algorithms for this little-explored approach to sparse SVMs [73].

3.4.2 Difference of Convex functions (DC) Programming

As mentioned previously, L_q-SVM has a non-convex objective function when $q \in (0, 1)$. Therefore standard optimization routines may fail to find good minima of the objective. We address this difficulty by adopting the technique of DC programming, which seeks to decompose an objective function into convex parts, for which global minima can be found, and then obtain a simpler non-convex function of these parts, for which we are more likely to find good minima. In this section, we briefly review DC programming technique and remarks on its convergence properties.

Given a non-convex objective function J and a DC decomposition of J,

$$J(\cdot) = G(\cdot) - H(\cdot) \quad , \tag{21}$$

the DC algorithm for minimizing J can be described in Table 5,

Table 5: General DC Algorithm Framework

Set an initial estimation $w^0 \in \mathrm{Dom}(J) = \{w \in R^d : J(w) < \infty\}$
$t = 0$
Repeat
Select $\beta^t \in \partial H(w^t)$ arbitrarily
Select $w^{t+1} \in \partial G^*(\beta^t)$ arbitrarily
$t = t + 1$
Until convergence, e.g. $\min_j(\mid w_j^t - w_j^{t-1} \mid) \leq tol$

where $G, H : R^d \to R$ are lower semi-continuous, proper convex functions, $G^*(\beta) = \sup\{w^T\beta - G(w)\}$ is the conjugate function of $G(w)$, and $\partial H(w)$, $\partial G^*(\beta)$ are subdifferentials of $H(w)$ and $G^*(\beta)$, respectively.

The subdifferentials of a lower semi-continous, proper convex function $F(w)$, can be defined as,

$$\partial F(w) = \{\beta \in R^d \mid F(w + \triangle w) \geq F(w) + (\triangle w)^T \beta \ , \ \forall \triangle w \in R^d\} \quad . \tag{22}$$

If $F(w)$ is differentiable, then $\partial F(w) = \{\nabla F(w)\}$. Furthermore, using standard results on convex optimization [90] [Theorem 23.5], the subdifferentials can be computed as in Eqn. (23). Therefore, a more analytical specification of the DC algorithm framework can be obtained (see Table 6).

$$\begin{array}{rcl} \partial F(w) & = & \arg\max_{\beta \in R^d}\{w^T\beta - F^*(\beta)\} \\ \partial F^*(\beta) & = & \arg\max_{w \in R^d}\{\beta^T w - F(w)\} \end{array} \tag{23}$$

Table 6: DC Algorithm Framework Implementation

Set an initial estimation $w^0 \in \text{Dom}(J) = \{w \in R^d : J(w) < \infty\}$
$t = 0$
Repeat
Select $\beta^t \in \partial H(w^t)$ arbitrarily
Select $w^{t+1} \in \arg\min_{w \in R^d}\{G(w) - w^T\beta^t\}$ arbitrarily
$t = t + 1$
Until convergence, e.g. $\min_j(\mid w_j^t - w_j^{t-1} \mid) \leq tol$

Pham Dinh and Le Thi (1998) proved that DC algorithm converges to a local minimum, controlled by the initialization value w^0 and the DC decomposition of the objective function [85].

Theorem 1 [85] Given a nonconvex objective function J and a DC decomposition $J = G - H$, if $\text{Dom}(G) \subset \text{Dom}(H)$ and $\text{Dom}(H^*) \subset \text{Dom}(G^*)$, then it holds for the DC algorithm that

(i) Sequences $\{w^t\}_{t \in N}$ (primal), $\{\beta^t\}_{t \in N}$ (dual) are well defined.

(ii) Objective value sequences $\{G(w^t) - H(w^t)\}_{t \in N}$ (primal), $\{H^*(\beta^t) - G^*(\beta^t)\}_{t \in N}$ (dual) are monotonously decreasing, respectively.

(iii) If the minimum of J is finite and the sequence $\{w^t\}_{t \in N}$ is bounded, every limit point $\tilde{w}$ of the sequence is a critical point of J, that is the point satisfying the local optimality condition of J. In particular, if $J(w^{t+1}) = J(w^t)$, then w^t is a critical point of J in Eqn. (21).

Pham Dinh and Le Thi (2008) further extend the convergence analyses to objective functions that are defined over convex sets [68]. Let $\Omega \subset R^d$ be a nonempty closed convex set, then the optimization problem

$$\min_{w \in \Omega} J(w) = G(w) - H(w) \tag{24}$$

can be transformed into unconstrained DC programming problem

$$\min J(w) = \tilde{G}(w) - H(w) \quad , \tag{25}$$

where $\tilde{G}(w) = G(w) + \Gamma_\Omega(w)$ is a semi-continuous proper convex function with indicator function $\Gamma_\Omega(w) = 0$, if $w \in \Omega$ and $\Gamma_\Omega(w) = +\infty$ otherwise.

Moreover, DC programming can also be applied to functions defined on R^{2d}_+, since the dual cone of R^{2d}_+ is still R^{2d}_+, all the derivations of DC algorithm would follow similarly [38]. Note that unlike convex concave procedure (CCCP) [119] and surrogate maximization / minimization approaches [124], DC programming is capable of optimizing non-convex objective functions that are non-smooth.

These convergence guarantees make the DC programming suitable for optimizing the fractional-norm SVM, that is L_q-SVM with $q \in (0, 1)$.

3.4.3 Solving the Fractional-norm SVM

The crucial point in applying the DC algorithm framework is to define a proper DC composition of the objective function. We propose a DC decompositions for the

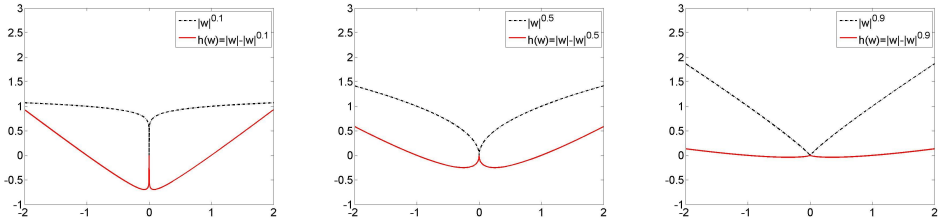
fractional-norm SVM problem under local linear approximation. In this section, we present the proposed DC decompositions, as well as the details of the DC iteration of the nested algorithm.

3.4.3.1 DC Decomposition under Local Linear Approximation

If we approximate $L_{q\in(0,1)}$-penalty locally with L_1-penalty, we can formulate a decomposition for the fractional-norm SVM problem as follows:

$$\begin{aligned} G(w) &= C\sum_{i=1}^{n} L\{f(x_i), y_i\} + \sum_{j=1}^{d} |w_j| \\ H(w) &= \sum_{j=1}^{d} (|w_j| - |w_j|^q) \end{aligned} \quad , \tag{26}$$

where decision function $f(x_i) = \sum_j w_j x_{ij} - \gamma$, $L\{f(x_i, y_i\} = [1 - y_i f(x_i)]_+$ is the hinge loss function$G(w)$ is a L_1-SVM type problem, and $H(w)$ measures the difference between the linear approximation and the L_q-penalty itself.



(a) q=0.1 (b) q=0.5 (c) q=0.9

Figure 10: Illustrations of the DC Decomposition of Fractional-norm SVM

If we define function $h(w_j) =\mid w_j \mid - \mid w_j \mid^q$, as per the graphical illustrations of the decompositions (depicted in Figure 10), we can see that the functions h are not convex on R. However, the restrictions on R_+ are convex. Therefore, we rewrite the fractional-norm SVM problem (Eqn. (20)) into

$$\begin{aligned} \min\ J(w^+, w^-, \gamma) = \quad & C\sum_{i=1}^{n} L\{f(x_i), y_i\} + \sum_{j=1}^{d} (w_j^+ + w_j^-)^q \\ \text{s.t.} \quad & w^+, w^- \in R_+^d, \gamma \in R \end{aligned} \quad , \tag{27}$$

where decision function $f(x_i) = \sum_j (w_j^+ - w_j^-) * x_{ij} - \gamma$.

And we consider the following DC decomposition for Eqn. (27):

$$\begin{array}{l} G(w^+, w^-, \gamma) = C\sum_{i=1}^{n} L\{f(x_i), y_i\} + \sum_{j=1}^{d}(w_j^+ + w_j^-) \\ H(w^+, w^-, \gamma) = \sum_{j=1}^{d}[(w_j^+ + w_j^-) - (w_j^+ + w_j^-)^q] \end{array} \quad . \tag{28}$$

Therefore, the DC algorithm is still applicable to solve the fractional-norm SVM problem given the convergence guarantees stated in Section 3.4.2.

3.4.3.2 DC Algorithm Specification for Fractional-norm SVM

According to the proposed DC decomposition (Eqn. (28)), the iterative optimization scheme for Problem (27) can be specified as in Table 7.

Table 7: DC Algorithm for Fractional-norm SVM

Set an initial estimation $w^0 \in \mathrm{Dom}(J)$
$t = 0$
Repeat
Determine w^{t+1} by optimizing
$\min_{w^+, w^- \in R_+^d} \quad C\sum_i L\{f(x_i), y_i\} + \sum_j (1 - \lvert\beta_j^t\rvert)(w_j^+ + w_j^-)$
where $\beta_j^t = \partial h(w_j^t)$
$t = t + 1$
Until convergence of w

At the t-th DC iteration, we first determine $((\beta^t)^+, (\beta^t)^-) \in \partial H((w^t)^+, (w^t)^-)$, where

$$\partial H(w_j) = \mathrm{sign}(w_j)[1 - \frac{q}{((w_j)^+ + (w_j)^-)^{1-q}}] \ , \quad w_j \neq 0 \quad . \tag{29}$$

To handle the computational issue when $w_j^t \approx 0$, we can add a σ term to the denominator [19, 93, 38]. Next, we determine $((w^{t+1})^+, (w^{t+1})^-) \in \partial G^*((\beta^t)^+, (\beta^t)^-)$. According to Eqn. (23), we need to solve the following optimization problem,

$$\arg \min_{w^+, w^- \in R_+^d} C\sum_i L\{f(x_i), y_i\} + \sum_j [1 - ((\beta_j^t)^+ + (\beta_j^t)^-)](w_j^+ + w_j^-) \quad . \tag{30}$$

Problem 30 can be further reformulated to a linear programming problem,

$$\min_{w^+,w^-\in R^d_+} C[e_n - YX(w^+ - w^-) + \gamma Y e_n]_+ + \varrho^T(w^+ + w^-) \quad , \tag{31}$$

where data matrix $X = [x_1, \cdots, x_n]^T$, label matrix $Y = \text{diag}\{\ y_1, \cdots, y_n\}$, $e_n = [1, \cdots, 1]^T \in R^n$, $\varrho_j = 1 - |\beta_j^t|$ and $\varrho = [\varrho_1, \cdots, \varrho_d]^T \in R^d_+$.

Equation (31) is a reweighted L_1-SVM problem. Recently, several efficient algorithms have been proposed for optimizing L_1-SVM [128, 37, 74]. Among all these methods, we extend the work of [74], which is very competitive, to solve our reweighted L_1-SVM problem. As an observation from [74], we can show that the optimal solution to the corresponding exterior penalty problem of Problem (31),

$$\begin{array}{ll} \min & -\epsilon e_n^T\nu + \frac{1}{2}(||\ (X^TY\nu - \varrho)_+\ ||^2 + ||\ (-X^TY\nu - \varrho)_+\ ||^2 \\ & +(e_n^TY\nu)^2 + ||\ (\nu - Ce_n)_+\ ||^2 + ||\ (-\nu)_+\ ||^2) \end{array} \quad , \tag{32}$$

is also the solution of Problem (31), where $\nu \in R^n$ corresponds to the unconstrained version of the Lagrange multipliers of Problem (31).

Problem (32) is an unconstrained optimization problem, and could be solved using Newton method, which is a second order algorithm but enjoys fast convergence as shown in our experiments. We compute the optimal solution of Problem (31) at each DC iteration as:

$$\begin{array}{rcl} w & = & \frac{1}{\epsilon}[(X^TY\nu - \varrho)_+ - (X^TY\nu - \varrho)_+] \\ \gamma & = & -\frac{1}{\epsilon}e_n^TY\nu \end{array} \quad .$$

We initialize the DC optimization with the optimal solution of the corresponding L_1-SVM, i.e. $w^0 \in \arg\min C\sum_i L\{f(x_i), y_i\} + \sum_j |w_j|$. And we terminate the DC optimization whenever $||\ w^{t+1} - w^t\ ||_\infty \leq \tau$ ($\tau = 10^{-3}$ for example) or the maximal number of DC iterations is reached.

3.4.4 Method Immplementation

In our experiments, all the methods are implemented in MATLAB. We use LibSVM package [24] for the standard SVM optimization. And we use the L_1-SVM optimization code from Fung and Mangasarian (2004).

For fractional-norm SVM optimization, we initialize the DC optimization with the solution of the corresponding L_1-SVM problem (that is $\varrho = 1_d$). We modify the L_1-SVM code to solve the reweighted L_1-SVM problem at each DC iteration. We adopt an annealing strategy to adjust σ (used to avoid numerical issue when computing subdifferentials of H) value. The algorithm starts from $\sigma = 10^{-3}$, optimizes the problem through DC algorithm until convergence, then scales down σ by a factor of 10 and reruns the DC algorithm initialized with the optimization result of the previous iteration. We stop the annealing procedure when $\sigma = 10^{-10}$ or maximum number of DC iterations is reached.

For R2W2 optimization [113], the W^2 term in the objective function is minimized using LibSVM package and the R^2 term is minimized using buildin MATLAB routine *quadprog*. At each iteration, 10% of the remaining features with the smallest nonzero scaling variables are removed. While for low-dimensional data sets such as *wpbc24, wpbc60*, and the synthetic data sets, only the feature with the smallest nonzero scaling variable is discarded at each R2W2 iteration.

For DC-FSV optimization [67], we solve the linear programming problem at each DC iteration using buildin MATLAB routine (*linprog*). In the final output, the weight elements with small relative magnitude, i.e. $\frac{|w_j|}{\max_k(|w_k|)} < 10^{-4}$, are set to zero. However, it's easily to show that the linear programming problem $\arg\min_w\{G1(w,\gamma,y,z) - w^T(\partial H1(w))\}$ could become unbounded during the DC optimization. Whenever this situation occurs, we stop the optimization and use the result from the previous DC iteration for thresholding and performance computation. In contrast, our fractional-norm SVM method doesn't have this computational issue since it guarantees $\varrho \geq 0$

(see Eqn. (31)) during the DC optimization.

For SCAD-SVM optimization, following the experiments in [34, 120], we use the standard SVM solution for initialization, and set the SCAD penalty parameter $a = 3.7$ and the thresholding value for removing features as 10^{-3}.

For LQA optimization, we use the linear discrimination analysis solution for initialization [73]. At each LQA iteration, we discard features with small weight values, $\mid w_j \mid < 10^{-3}$. Since its optimization formulation don't guarantee a positive-definite Hessian, we thus use Matlab routine *pinv*, which computes the Moore-Penrose pseudoinverse [9] of the hessian matrix, to get an approximation solution of the unconstrained quadratic programming problem at each LQA iteration.

3.4.4.1 Tuning Parameters

We employ the grid search procedure for parameter tuning. Note that for standard SVM and R2W2 method, we need to tune parameter $C > 0$ for the performance evaluation; for DC-FSV method, the tuning parameter is $\lambda \in (0, 1)$; for SCAD-SVM method, the tuning parameter is $\lambda \in (0, 1]$; for LQA method, the tuning parameters are $q \in (0, 2]$ and $C > 0$; for L_1-SVM method, the tuning parameters are $C, \delta > 0$; and for our fractional-norm SVM method, the tuning parameters are $C, \delta > 0$, and $q \in (0, 1)$. The candidate values of the parameters used for the experiments were

- $C \in \{2^{-7}, \cdots, 2^{-1}, 1, 2^1, \cdots, 2^7\}$,
- $\delta \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\}$,
- $\lambda^{DC-FSV} \in \{0.001, 0.002, \cdots, 0.004, 0.05, 0.1, 0.2, \cdots, 0.5\}$,
- $\lambda^{SCAD-SVM} \in \{0.1, 0.2, \cdots, 0.9, 1\}$,
- $q^{LQA} \in \{0.1, 0.2, \cdots, 1.9, 2\}$
- $q^{L_qSVM} \in \{0.1, 0.2, \cdots, 0.8, 0.9\}$.

3.4.5 Simulation Study

In this simulation study, our objective was to assess the ability of the algorithms to select a small number of target features in the presence of irrelevant and redundant features. We generate the synthetic datasets following the example used in [113]. There are $d = 202$ features and only the first six dimensions are relevant. The first three features x_1, x_2, x_3 were drawn as $x_j = yN(j, 1)$ and the second three features x_4, x_5, x_6 were drawn as $x_j = N(0, 1)$ with a probability of 0.7, otherwise the first three were drawn as $x_i = N(0, 1)$ and the second three as $x_j = yN(j - 3, 1)$. The remaining features $x_j, j = 7, \cdots, 202$ are noise and independently generated from $N(0, 20)$. The probability of label $y = 1$ or -1 is equal. We consider various training sample size: $n = 60, 70, 80, 90, 100, 110, 120$.

For each training sample size n, we repeat the following evaluation procedure 30 times for each method. We first generate a training data set of size n, and a testing data set of size 500. Then we evaluate the method's performance on the training/testing datasets. The final classification and feature selection performance of each method is computed as the average testing error and average number of selected features over the 30 runs.

We employ a grid search procedure for choosing the right parameters for each sample size. Candidate parameters for each method are listed in Section 3.4.4.1. For each method under each parameter setting, we perform the above evaluation procedure, and the score for this parameter setting is the averaged training error rate over the 30 runs. Then we select the parameter setting with the best score (ties are broken by choosing the sparser solutions).

We plot the trend of the classification performance of each method to the size of the training sample size in Figure 11. x axis is the number of the training sample size of the synthetic data sets. y axis is the average testing error rate of each method over the 30 replicas. The SVM method performs poorly on these synthetic data

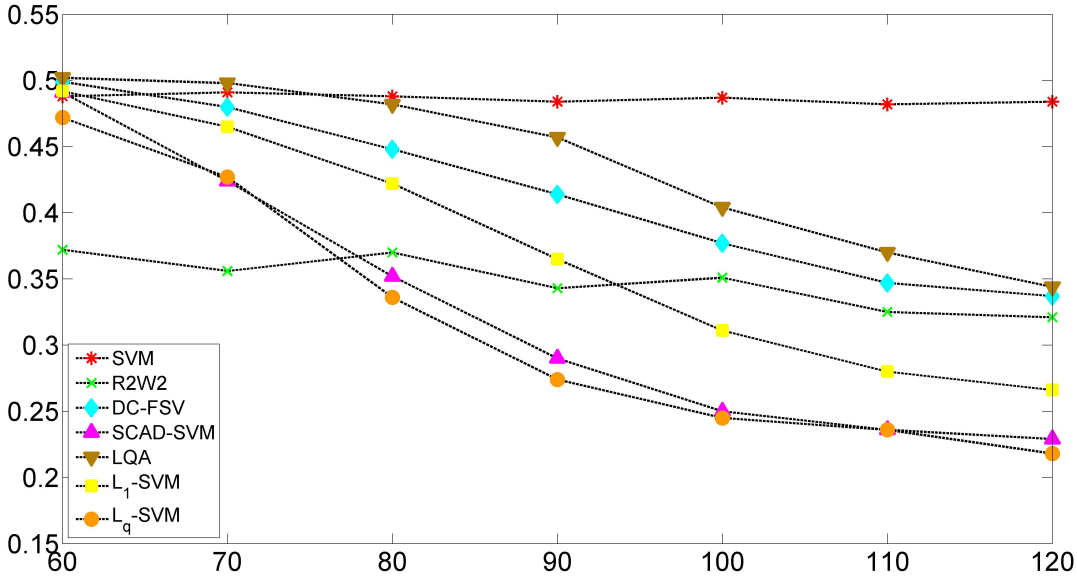


Figure 11: Classification Performance on Synthetic Data Sets

sets, obtaining near random prediction performance (0.5 testing error rate) in most cases. While testing error rates of the sparse SVM methods generally decreases as the training sample size n increases. This suggests that feature selection is necessary when many noise variables are present in the data set. The testing error rates of the L_q-SVM and SCAD-SVM methods decrease prominently compared to those of the LQA, DC-FSV and L_1-SVM methods. R2W2 method has the best prediction performance on synthetic data sets with training sample size $n = 60, 70$. The L_q-SVM and SCAD-SVM methods surpass R2W2 method on synthetic data sets with training sample size $n = 80, 90, 100, 110, 120$. Furthermore, our L_q-SVM methods slightly performs better than SCAD-SVM method in all cases.

Table 8 lists the average number of features selected by each method over the 30 replicas. Since the SVM method is not designed to select variable, we didn't include it in this comparison. It is observed that our fractional-norm SVM method consistently selects the smallest feature subsets in all cases. The DC-FSV, LQA method doesn't perform well on these synthetic data sets. In addition, we found that several sparse SVM methods such as R2W2, SCAD-SVM and L_q-SVM methods are all able to

include the relevant features into their feature selection results in most of the runs. Our fractional-norm SVM method tends to be less affected by the irrelevant features and thus achieves the sparsest feature selection results among these methods.

Table 8: Feature Selection Performance (Number of Selected Features) on Synthetic Data Sets

	R2W2	DC-FSV	SCAD-SVM	LQA	L_1-SVM	L_q-SVM
n=60	12	49	53	17	27	**11**
n=70	14	57	50	14	31	**12**
n=80	17	63	45	16	32	**12**
n=90	18	70	40	25	32	**12**
n=100	19	74	38	27	31	**12**
n=110	21	78	39	27	31	**12**
n=120	24	83	44	28	32	**13**

3.4.6 Empirical Study on Real World Data Sets

We further conduct a comparison study on seven real-world data sets (see Table 9) through 10-fold cross validation. For each data set, we permuted the order of the data samples before 10-fold partition, and we partitioned the data samples such that each fold has the same number of negative data samples and the same number of positive data samples. The performance of each method is measured in terms of averaged testing error rate and sparsity of the 10-fold cross validation under their respective optimal parameter settings, and its average computation time of a 10-fold cross-validation averaged over all the parameter settings. In which, *sparsity* is defined as the ratio of size of selected feature subset and the total feature size.

We employ the following grid search procedure for choosing the right parameter setting for each data set. For each possible parameter setting (see Section 3.4.4.1), we perform the 10-fold cross-validation, and the score for this parameter setting is the averaged training error rate over the 10-fold cross-validation. The parameter setting with the best score (ties are broken by choosing the sparser solutions) is chosen for

the final comparison.

Table 9: Statistics of the Data Sets

Data sets	Feature Size (d)	Sample Size (n)	Class Distribution (n_+/n_-)
arcene	10000	200	88 / 112
ColonCancer	2000	62	22 / 40
OvarianCancer	592	72	35 / 37
PancreaticCancer	6771	181	80 / 101
ProstateCancer	12600	102	52 / 50
wbpc24	32	155	28 / 127
wbpc60	32	110	41 / 69

3.4.6.1 Data Set Description

We use four public microarray gene expression data sets and three mass spectrometry data sets in the comparison study (see Table 9).

- *arcene* data set is from NIPS 2003 feature selection challenge [47]. The task of Arcene is to distinguish cancer (ovarian or prostate cancer) versus normal (healthy or control) patterns from mass spectrometric data. Note that in our experiments, we combine the training data (100 cases) and validation data (100 cases) of the *arcene* data set.

- Colon cancer data set [2] consists of 62 tissue samples (22 normal and 40 colon cancer) probed by oligonucleotide arrays. The data set contains expression values of the 2000 genes with highest minimal intensity across the 62 tissues. Some genes are non-human genes.

- The ovarian cancer data set [44] consists of mass spectrometry metabolomic profiles of 37 ovarian cancer patients and 35 benign controls. Each metabolomic

profile contains intensity values of the same 592 multimode features (360 in posion-mode and 232 in neg-ion-mode) extracted from the liquid chromatography/time-of-flight mass spectra of the patient serum.

- Hingorani et al. (2003) [51] explored the ability of the low molecular weight information in discriminating premalignant pancreatic cancer compared to control animals. A data set consists of 80 PANIN (pancreatic intraepithelial neoplasias) and 101 control murine sera samples were provided from their website.

- The prostate cancer data set [98] consists of 52 prostate tumors and 50 nontumor prostate samples. The gene expressions were measured using high-density oligonucleotide arrays with probes for 12600 human genes and ESTs.

- The *wpbc24* and *wpbc60* data sets are build on the the Wisconsin prognostic breast cancer patient follow up data set (*wpbc*) [78], including the 30 nuclear features plus diameters of excised tumor and number of positive lymph nodes. For *wpbc24* data set, the two classes were patients with recurrence before 24 months (28 cases), and patients with recurrence after 24 months (127 cases). For *wpbc60* data set, the two classes were patients with recurrence before/after 60 months (41 cases and 69 cases, respectively).

3.4.6.2 Performance Evaluation

The first set of experiments evaluated the seven methods on the *ColonCancer*, *OvarianCancer*, *wpbc24*, *wpbc60* data sets. Table 10,11 depicts their performance over the 10-fold cross validation under the optimal parameter setting, respectively. The values in the parentheses are the standard errors over the 10-fold cross validation of the corresponding mean values.

Comparing to L_1-SVM method, many of the L_0-SVM approximation methods (DC-FSV, SCAD-SVM and our fractional-norm SVM methods) were able to obtain

Table 10: Classification Performance (Accuracy %)

Datasets	SVM	R2W2	DC-FSV	SCAD-SVM	LQA	L_1-SVM	L_q-SVM
ColonCancer	91	**82.6**	81.9	77.4	77.4	75.5	82.1
	(4.6)	(11.1)	(14.6)	(11.5)	(11.5)	(14.5)	(14.1)
OvarianCancer	78.7	48.4	73.0	70.4	58.8	72.9	**76.1**
	(14)	(18.2)	(20.8)	(19.9)	(11.3)	(20.7)	(17.8)
wbpc24	78.8	80.1	65.4	80.1	80.8	78.8	**81.3**
	(5)	(4.4)	(12.7)	(4.4)	(4.7)	(7.6)	(5.4)
wbpc60	66.2	65.4	63.5	67.2	68.0	61.0	**68.0**
	(8)	(8.5)	(9.7)	(10.8)	(10.0)	(10.9)	(10.3)
on-average	78.7	69.1	71.0	73.8	71.2	72.0	**76.9**

Table 11: Feature Selection Performance (Sparsity)

Datasets	R2W2	DC-FSV	SCAD-SVM	LQA	L_1-SVM	L_q-SVM
ColonCancer	0.005	**0.003**	**0.003**	0.005	0.011	0.005
	(0.0007)	(0.0015)	(0.0008)	(0.0009)	(0.0016)	(0.0007)
OvarianCancer	0.068	0.063	0.053	**0.026**	0.071	0.068
	(0.016)	(0.005)	(0.005)	(0.086)	(0.005)	(0.006)
wbpc24	0.869	0.481	**0.091**	0.791	0.781	0.119
	(0.0)	(0.0)	(0.1)	(0.1)	(0.033)	(0.038)
wbpc60	0.691	0.491	**0.184**	0.228	0.788	0.306
	(0.16)	(0.044)	(0.083)	(0.026)	(0.041)	(0.041)
on-average	0.408	0.259	**0.083**	0.263	0.413	0.125

sparser representation of these data sets while archive improved or at least similar prediction performance. This observation support the analysis that L_0-SVM optimization is able to achieve more aggressive feature selection than L_1-SVM optimization. Overall, our fractional-norm SVM method shows robust performance: consistently achieving improved prediction performance as well as sparse solutions with decent runtime.

Table 12: CPU Runtime (Seconds) of a 10-fold Cross-Validation

Datasets	SVM	R2W2	DC-FSV	SCAD-SVM	LQA	L_1-SVM	L_q-SVM
ColonCancer	1	2014	207	867	921	4	21
OvarianCancer	1	9572	220	100	669	2	22
wbpc24	1	1583	15	36	5	2	14
wbpc60	1	1524	6	30	8	1	7
on-average	1	3673	112	258	401	2	16

We can see from Table 12 that Mangasarian's L_1-SVM optimization algorithm is very efficient although it is a second-order algorithm. The computation time of our fractional-norm SVM method is roughly bounded by that of the L_1-SVM algorithm times the maximum number of DC iterations. R2W2 method requires the largest computational time. There are three main reasons: i) The number of R2W2 iterations depends on the number of features. ii) Each R2W2 iteration includes optimization of two quadratic programming problems $\min R2$, $\min W2$, and a gradient descent update. iii) The optimization time of the two quadratic programming problems increases dramatically when the remaining feature size becomes too small (< 30). Thus we terminate R2W2 procedure if runtime of current iteration exceeds 100 seconds.

We also notice that the computation time of DC-FSV, SCAD-SVM, and LQA methods on *OvarianCancer* and *ColonCancer* data sets (feature size 592, 2000, respectively) increases a lot comparing to their computation time on *wpbc24* and *wpbc60* data sets (feature size 32). This fact suggests that these methods might not be scalable to high dimensional datasets. The same conclusion can also be derived from the

analysis on their optimization formulation. At each DC iteration, DC-FSV algorithm optimizes a linear programming problem of $O(n) + O(d)$ variables and $O(n) + O(d)$ constraints. At each optimization iteration, SCAD-SVM and LQA methods, need to compute the inverse or pseudo-inverse of its hessian matrix, which is a $d \times d$ matrix. Therefore, if thresholding on the initial solution (e.g. solution of standard SVM) wasn't able to reduce the feature size substantially, the three methods could fail on high-dimensional data sets .

Table 13: Classification Performance (Accuracy) on High Dimensional Data Sets

Datasets	SVM	R2W2	L_1-SVM	L_q-SVM
arcene	91 (4.6)	64 (15.1)	58 (12.3)	**70** (12.7)
PancreaticCancer	70.2 (10.4)	61.3 (12.4)	53.6 (9.2)	**67.4** (11.4)
ProstateCancer	91.2 (5.7)	88.4 (8.9)	85.2 (12.8)	**91.2** (5.7)
on-average	84.1	71.2	65.6	**76.2**

In the second set of experiments, we compare the methods on the three high-dimensional data sets in terms of feature selection performance (Table 13), prediction performance (Table 14), and computational cost (Table 15). DC-FSV, SCAD-SVM, and LQA methods are not included in this set of experiments, because we keep getting out-of-memory error messages when applying these methods onto these high-dimensional data sets. Overall, the comparison showed that, our fractional-norm SVM method is able to achieve sparse solutions, comparable to those of the R2W2 method, with better prediction performance with reasonable running time.

We can see from Table 15 that SVM learning time increases with feature dimensions as well as sample size. R2W2 method still require the largest computational time. We apply the same termination rule for R2W2 method, the method will be terminated if the running time of current iteration exceeds 100 seconds. For L_1-SVM

Table 14: Feature Selection Performance (Sparsity) on High Dimensional Data Sets

Datasets	R2W2	L_1-SVM	L_q-SVM
arcene	**0.003** (0.0005)	0.004 (0.003)	**0.003** (0.0002)
PancreaticCancer	**0.005** (0.0006)	0.007 (0.0003)	0.007 (0.0005)
ProstateCancer	**0.001** (0.0002)	0.002 (0.0002)	**0.001** (0.0002)
on-average	**0.003**	0.004	0.004

method, the complexity at each Newton iteration is of $O(n^2d) + O(n^3)$, thus we also see an increase of computational time on these high-dimensional data sets. But it is still very efficient comparing to that of the R2W2 method. The computation time of our fractional-norm SVM method is still roughly bounded by that of the L_1-SVM algorithm times the maximum number of DC iterations.

Table 15: CPU Runtime (Seconds) of a 10-fold Cross-Validation

Datasets	SVM	R2W2	L_1-SVM	L_q-SVM
arcene	144	6075	81	875
PancreaticCancer	25	6769	98	592
ProstateCancer	10	1192	71	456
on-average	60	4679	58	641

In summary, we observed that on data sets such as *wpbc24*, *wpbc60* and *ProstateCancer*, the prediction performance of feature-selecting SVM methods such as fractional-norm SVM, SCAD-SVM are better than that of the standard SVM method, indicating that many features in these data sets may be irrelevant. In data sets like *OvarianCancer*, *ColonCancer* and *PancreaticCancer*, the prediction error rate remains roughly the same after applying feature-selecting SVM methods such as R2W2 or fractional-norm SVM, suggesting that these data sets may contain redundant features. In *arcene* data set, none of the sparse SVM methods beats the standard SVM

method in prediction accuracy. This is probably because *arcene* data set contains both ovarian cancer or prostate cancer patients, and these two types of cancer should correspond to different subset of biomarker candidates. The random splitting in the 10-fold cross validation experiment probably did not generate balanced partitions of ovarian/prostate cancer patients. The different ovarian/prostate cancer patient distribution in the training/testing partition thus resulted in bad performance for these sparse SVM classifiers. In all three cases, our fractional-norm SVM method can robustly realize a proper trade-off between prediction accuracy and the number of selected features within decent runtime. Moreover, the sparse decision functions learned with our method are generally more predictive than those produced by L_1-SVM or other L_0-SVM approximation methods.

3.5 Discussion

In order to assist high throughput biomarker discovery, we investigate feature selecting support vector machine based on convex relaxations of L_0-SVM formulation. We propose the mixed-integer support vector machine and explore the mixed-integer nonlinear programming techniques which have not previous been widely used in machine learning. We further propose an practical solution to the fractional-norm SVM problem, an more efficient L_0-SVM approximation, using the difference of convex functions programming technique. The empirical study support the effectiveness of the fractional-norm SVM over other commonly-used sparse SVM methods.

CHAPTER IV

LEARNING PROTEIN FOLDING ENERGY FUNCTION

4.1 Introduction

Proteins are polymers assembled from 20 naturally occurring amino acids, which fold to a unique, biologically active, three-dimensional conformation called the *native structure*. Their functions are governed by their three-dimensional structures, which in turn are fully determined by their amino acid sequences. Predicting the native structure of a protein from its amino acid sequence, is one of the most important and challenging scientific problems in contemporary biology and chemistry [36], [100]. The capability to reliably make such predictions would allow biochemists to design drugs more efficiently, understand various biological processes in details, and answer many fundamental questions about biological systems, diseases, immune response, and more.

The experimental determination of protein tertiary structure is a time consuming and expensive process. Hence, computational methods play an essential role in the native structure prediction of proteins. There are three classes of computational based protein structure prediction approaches: homology modeling, threading and *ab initio* folding. Both homology modeling and threading methods suffer from the fundamental limitation that the query protein sequence must be evolutionarily related to some proteins with known tertiary structure [99]. For a query protein, if none of its homologous sequences has an experimentally resolved structure, the only remaining approach to predict its native structure is *ab initio* folding.

Ab initio folding attempts to find the native structure of a protein "from scratch" (see Figure 12). The fundamental assumption in *ab initio* folding is the existence of a

free energy function that assigns an energy value to each three-dimensional structure the protein can in principle assume. The native structure is assumed to be the one with the lowest energy. Although there are notable exceptions, this assumption holds true for the vast majority of the protein native structures. Thus, there are two main ingredients in *ab initio* folding: The design of a reliable energy function, and the development of an efficient approach to the search the space of all possible conformations for the one with the lowest energy. In this work, we focus on the first problem, namely, finding an energy function which allows for efficient and accurate determination of the native structure of proteins in an *ab initio* folding approach.

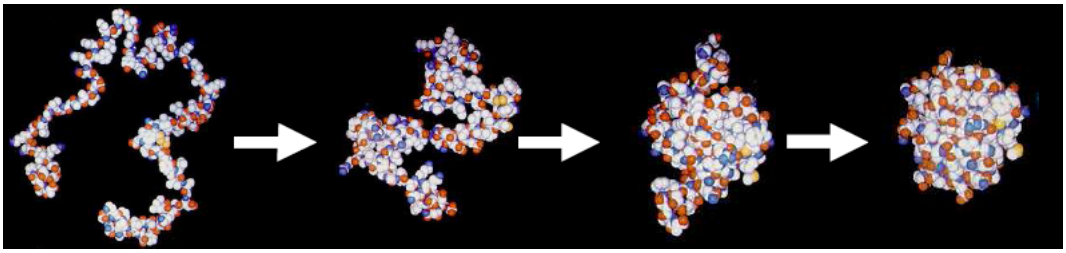


Figure 12: Illustration of *Ab initio* Folding

The energy functions used in *ab initio* folding are physics-based: for a given three dimensional configuration of a protein, one first calculates various terms contributing to the total energy such as electrostatic energy, covalent bonding energy, Van der Waals energy, etc., and then adds these terms to obtain the total energy. While these terms are based on physics, their functional forms are sometimes approximate, and the coefficients that appear are obtained by various fitting procedures. In this work, we represent the total energy of a configuration as a linear combination of these physics-based energy terms, and optimize the coefficients.

For a given protein, we represent the total energy of a given, candidate three-dimensional structure (conformation) s as $f(s) = w^T x$, where $x \in R^n$ represents the collection of the energy terms for the conformation s, and $w \in R^n$ is the weight coefficients. The fitness of a given energy function can be visually inspected by plotting

the total energy versus the structural dissimilarity to the native structure. For a given protein with known *native structure*, one generates many possible conformations, and obtains the total energy and dissimilarity from the native structure for each. There are various notions of structural similarity used in the literature, such as the root mean squared distances (RMSD) [62] between the building blocks (e.g., atoms) of the protein as represented in two candidate structures aligned in three-dimensional space.

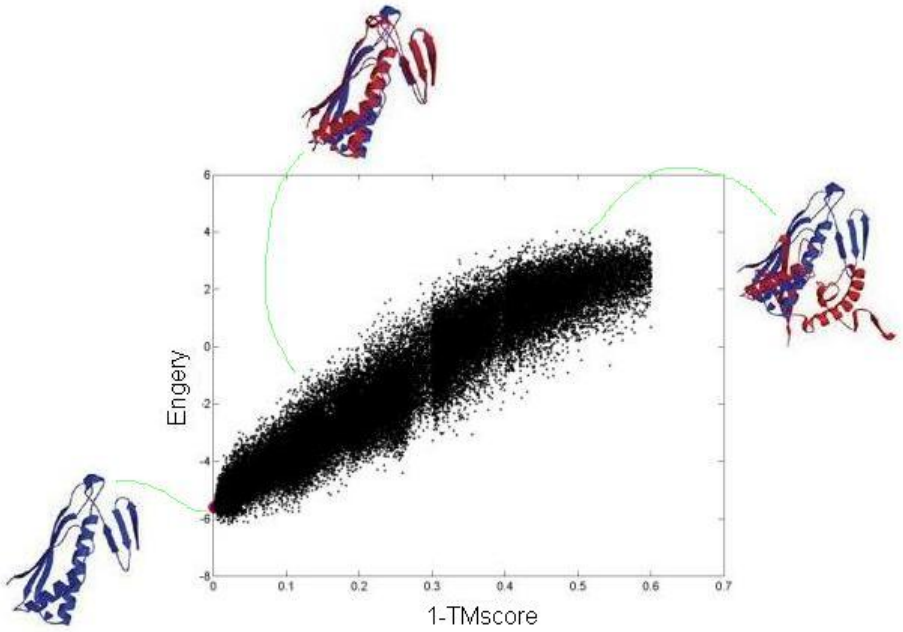


Figure 13: Energy versus Structural Dissimilarity Plot

Figure 13 shows such a plot for a desirable energy function. Each black dot represents a non-native conformation. As can be seen, the energy value is higher for conformations that have large dissimilarities from the native structure, with a roughly monotonic trend. This trend should be reproduced by any *ab initio* folding procedure ,which starts with some random conformation and searches the space of all possible conformations for the one with the lowest energy. Due to the monotonic trend, reducing the energy corresponds to getting closer to the native structure during *ab initio* folding procedure. If one can construct an energy function that has energy vs. dissimilarity plots like that of Fig. 13, one can hope to reproduce a similar trend for

proteins with unknown structure. As described below, we investigate the optimization problem by suitable training and test sets of proteins and cross-validation.

Let x_j be the vector of energy terms for the j-th conformation s_j of a protein, r_j be the structural dissimilarity between s_j and the native structure s_0 , and $E_j = w^T x_j$ be the energy of s_j. Treating the weight vector w as the unknown, the task of learning an *ab initio* protein folding energy function becomes a weight optimization problem. Much of the literature on this problem is based on maximizing the correlation (or related quantities) between the total energy and the dissimilarity. In this paper, we propose a ranking-based approach to this problem. Namely, given m conformations for each of a set of proteins, we propose to search for the weight vector w such that for each protein, a meaningful *subset* of the constraints below are satisfied.

- Total energy of the native structure is the minimum, that is, $E_0 < E_j$ for all $j = 1, \cdots, m$.
- Energy of random conformations with smaller structural dissimilarity are smaller than those with larger dissimilarity, that is, if $r_j < r_i$, then $E_j < E_i$.

This chapter is organized as follows. We begin in Section 2 by converting the weight optimization problem into a learning-to-rank task and then describe RankingSVM, a ranking-via-classification method that we utilize. Due to physicality constraints, we restrict the problem to non-negative weights. Two efficient algorithms to solve the constrained RankingSVM problem are described in Section 3. In Section 4, we analyze and discuss the experiment results.

4.2 Weight Learning By Ranking

The problem of learning protein energy function can be reduced into a learning-to-rank problem if we consider the ordering derived from the structure dissimilarity as the true ordering over the protein conformations, and the ordering derived from the energy

function as the predicted ordering. The corresponding learning-to-rank problem seeks to find a linear ranking function $f(s) = w^T x$ that optimally approximates the true ordering of the protein conformations. In our study, for any query protein, we expect the predicted ordering to satisfy the following requirements as many as possible:

i) Rank the native structure above other conformations

ii) Rank conformations with lower dissimilarities above those with higher dissimilarities

In this section, we review current machine learning approaches in learning-to-rank tasks, and then describe RankingSVM, which is the basis of our proposed methods.

4.2.1 Learning-to-Rank Methods

The simplest class of learning-to-rank methods is the pointwise approach. It can also be viewed as the ranking-via-regression approach. The most straightforward pointwise objective function is the MSE (mean squared error) between the rank/score of an object (e.g. a document, a protein conformation) in the true ordering and that in the predicted ordering, e.g. RankProp [21]. Another option would be to use criterion such as DCG (discounted cumulated gains) in the regression [30]. More sophisticated methods include: framework of ordinal regression, e.g. PRank [31], etc.

Because it is generally easier to obtain/model the preference over object pairs than the absolute rank order of objects, pairwise approach, also called as ranking-via-classification approach, was proposed. Typical methods include: RankingSVM, RankBoost, RankNet, etc. RankingSVM [50] seeks to maximize the Kendall τ statistics [64], which can be approximated by minimizing the number of mis-ordered object pairs in the predicted ordering. Rankboost [35] aims to minimize the weighted number of object pairs that are mis-ordered by the final ranking function. RankNet [17] is a probabilistic model for learning the pairwise preference.

Recently, listwise approach was proposed to directly tackle the ranking problem. In this class of methods, ranked list of objects are treated as instances in learning and the final ranking function is learned through the minimization of a listwise loss function. ListNet [20], SoftRank [105], SVM-MAP [118] are examples of methods in this category.

4.2.2 Ranking Via Support Vector Machines

Pointwise and pairwise approaches have the advantage that the existing theories and algorithms on regression and classification can be readily applied into the learning task. Moreover, pairwise approaches generally outperform the pointwise approaches and have been successfully applied to Information Retrieval applications [59],[56], [17]. Therefore, we adopt the ranking-via-classification approach to solve our problem: we first convert the data set by the difference space method [87], and then build a linear classifier on the converted data set. That is, given a data set $S = \{(x_i, r_i)\}_{j=0}^{m}$, we convert it into $S^{\text{diff}} = \{z_{ij} = x_i - x_j, y_{ij} = \text{Sign}(r_i - r_j)\}$, where z_{ij} is the pairwise difference vector, and y_{ij} is the sign of the rank difference of objects i, j. The optimal linear decision function learned on S^{diff} gives the optimal linear ranking function for the original dataset S.

RankingSVM method seeks to find a ranking function f that approximately maximizes the expected Kendall τ statistic: $\tau_S(f) = \frac{1}{N}\sum_{q=1}^{N}\tau(o_q^*, \hat{o}_q)$ on the training dataset $S = \{S_1, \cdots, S_N\}$, where data set $S_q = \{(x_i^q, r_i^q)\}_{i=0}^{m_q}$ contains the energy data of the qth protein, rank ordering o_q^* is the true ordering of the protein conformations derived from structure dissimilarity, and $\hat{o}_q$ is the approximated ordering determined by the linear ranking function $f(\cdot)$.

Kendall's τ statistic [64] for two finite orderings $o_q^*, \hat{o}_q$ is defined as

$$\tau(o_q^*, \hat{o}_q) = \frac{\text{concordant } \# - \text{discordant } \#}{\text{concordant } \# + \text{discordant } \#} \quad ,$$

where an object pair $s_i \neq s_j$ is called discordant if the orderings o_q^* and $\hat{o}_q$ do not agree in how they order s_i and s_j, and called concordant otherwise. In our study, discordant pairs are pairs of protein conformations s_i, s_j such that the learned ranking function $f(\cdot)$ ranks s_i above s_j, but the input preference ($r_j < r_i$) indicates s_j should have smaller energy than s_i; or vice versa.

For strict orderings on m instances, we have $\frac{m(m-1)}{2}$ = concordant #+discordant #. This will give too large a training set in our case, and we will work with a suitable subset below. A linear ranking function $f(x_i) = w^T x_i$ generates an ordering on the instances such that $s_i < s_j$ iff $f(s_i) < f(s_j)$. Maximizing the expected Kendall's τ statistic of a linear ranking function on the data set $S = \{(x_i, r_i)\}$ is equivalent to maximizing the pairwise agreement (concordant #). This optimization problem can be formulated as a search for the weight vector w that maximizes the number of inequalities of form $\mathrm{Sign}(\mathrm{r_i} - \mathrm{r_j})\mathrm{w^T}(\mathrm{x_i} - \mathrm{x_j}) \geq 1$ that hold true. It can be approximately solved by learning the SVM classifier [109] on the transformed data set, $S^{diff} = \{z_{ij} = x_i - x_j, y_{ij} = \mathrm{Sign}(r_i - r_j)\}$, where z_{ij} is the pairwise difference vector, y_{ij} is the sign of the rank difference of objects s_i, s_j, and ξ_{ij} are the slack variables.

$$\begin{aligned} \min \quad & \tfrac{1}{2} w^T w + \pi \textstyle\sum_{ij} \xi_{ij} \\ \text{s.t.} \quad & y_{ij} w^T z_{ij} \geq 1 - \xi_{ij}, \xi_{ij} > 0, \forall i, j \end{aligned} \tag{33}$$

4.3 Non-Negativity Constrained Weight Learning

The energy terms used in our optimization represent "costs", in the sense that the natural physical tendency of the protein is to decrease each one of these values. Each energy term, taken separately, represents a uniquely defined physical tendency. For the case of electrostatic interactions, two positive charges move away from each other in order to lower their interaction energy. Reversing the sign of this interaction energy would turn the repulsive force to an attractive one, hence resulting in an unphysical interaction. If we sacrifice the physicality of the energy function by picking negative

weights for some terms, it may be possible to obtain a better ranking on the collected set of conformations. Unfortunately, experience shows that such unphysical energy functions, while performing well on the chosen set of existing proteins, perform poorly when predicting new physical structures. This is partly because it is impossible to sample the whole set of possible conformations for a given protein, and the methods used to generate the conformations in the training set start from special, compact conformations that already satisfy various physicality properties. Dropping the positivity constraints could improve the ranking for these special conformations, but there will be very large, unsampled subsets of the set of possible conformations where the negative coefficients would result in incorrect foldings/orderings. Thus, one enforces a positivity constraint on the weights in order to avoid overfitting to the (small) set of sampled conformations.

We next describe two approaches to non-negative support vector machines (NNSVM).

4.3.1 Non-Negative L2-norm SVM

In this section, we propose a non-negative version of SVM by using an L_2 norm approach, and solve it through the exponential gradient (EG) algorithm [65].

Due to the characteristics of our problem, we formulate the optimization in primal form. Adding the non-negativity constraints to the standard SVM formulation gives the optimization problem,

$$\min_{w \geq 0} \frac{\nu}{2} w^T w + \frac{1}{l} 1_l^T (1_l - DAw)_+ \tag{34}$$

where $(u)_+ = \max(u, 0)$ sets the negative components of the vector u to zero, A denotes the data matrix with rows given by the z_{ij}s, $D = \mathrm{diag}(\mathrm{y}_1, \cdots, \mathrm{y}_\mathrm{l})$ is the label matrix, $1_l = [111 \ldots 1]^T$ is an l-dimensional vector of 1s, and l is the total number of data points (i.e. total number of pairwise difference vectors in our study).

The objective function in Eqn. (34) is non-differentiable, hence typical optimization methods cannot be directly applied to this problem. To address this issue, we use the L_2-norm of the hinge loss variables in the objective function. This type of SVM has gained popularity in large scale classification because the resulting objective function $J(w)$ is a piecewise quadratic and strongly convex function, and efficient algorithms such as coordinate descent [25] can be applied. The Non-Negative L_2-norm SVM (NNL2SVM) objective function is,

$$J(w) = \min_{w \geq 0} \frac{\nu}{2} w^T w + \frac{1}{2l} \mid\mid (1_l - DAw)_+ \mid\mid^2 \tag{35}$$

We use the exponential gradient (EG) algorithm [65] to solve this NNL2SVM problem because its optimization is naturally constrained to the non-negative space R^n_+. The algorithm is summarized in Table 16.

Table 16: EG Algorithm for NNL2SVM Problem

Initialize $w^0 = \frac{1}{n} 1_n$ so that $\mid\mid w^0 \mid\mid_1 = 1$
For $t = 0, 1, 2, \ldots$
Compute $\bigtriangledown J(w^t) = \nu w^t - \frac{1}{l} A^T D(1_l - DAw)_+$
For all $j = 1, \cdots, n$
Update $w_j^{t+1} = w_j^t e^{-\eta \bigtriangledown w_j^t}$
Normalize w^{t+1}

where the learning rate $\eta = 1/R$ with $R = \max_{ij}(\max_k z_{ij,k} - \min_k z_{ij,k})$, where $z_{ij,k}$ denotes the kth component of the feature vector $z_{ij} = x_i - x_j$. R is the largest value over the sample set of the maximum difference $\max_k z_{ij,k} - \min_k z_{ij,k}$ between the components of a feature vector z_{ij}.

The standard normalization sets $\mid\mid w^{t+1} \mid\mid_1 = 1$. We also investigate another normalization method that enforces $\mid\mid w^{t+1} \mid\mid_1 \leq \mid\mid w^t \mid\mid_1$ by keeping w^{t+1} unchanged if $\mid\mid w^{t+1} \mid\mid_1$ is less than $\mid\mid w^t \mid\mid_1$, and setting its norm to $\mid\mid w^t \mid\mid_1$ otherwise.

4.3.2 Non-Negative L1-norm SVM

Another approach to the NNSVM problem is to add non-negativity constraints to the L1-SVM formulation [13] and extend the existing L1SVM algorithm [37],[74] to solve the resulting NNL1SVM problem. The optimization problem is,

$$\begin{array}{lll} \min & 1_n^T w + \pi 1_l^T \xi & \\ & DAw & \geq \quad 1_l - \xi \\ \text{s.t.} & & \\ & w, \xi & \geq \quad 0 \end{array} \tag{36}$$

We solve this problem using an approach described in [74]. Proposition 1 in [74] states that for any $\epsilon \in (0, \bar{\epsilon}]$ for some $\bar{\epsilon} > 0$, the optimal solution of the exterior penalty problem gives an exact solution to the original, primal problem. The corresponding exterior penalty problem can be derived by assigning quadratic penalty terms to the constraints of the dual problem. The exterior penalty problem of Eqn. (36) minimize the following objective function,

$$J(\mu) = \quad -\epsilon 1_l^T \mu + \tfrac{1}{2}(|| \, (A^T D\mu - 1_n)_+ \, ||^2 + || \, (\mu - \pi 1_l)_+ \, ||^2 + || \, (-\mu)_+ \, ||^2) \, . \tag{37}$$

Problem 37 is an unconstrained optimization problem. We solve it using the generalized Newton method described in Table 17.

Table 17: Newton Method for NNL1SVM Problem

Initiate $t = 0$ and $\mu^1 = 1_l$
Repeat
$t = t + 1$
$\mu^{t+1} = \mu^t - \zeta_t(\delta I_l + \partial^2 J(\mu^t))^{-1} \bigtriangledown J(\mu^t)$
ζ_t is the largest number in $\{1, \frac{1}{2}, \frac{1}{4}, \cdots, \}$
such that $J(\mu^t) - J(\mu^t + \zeta^t d^t) \geq -\frac{\zeta^t}{4} \bigtriangledown J(\mu^t) d^t$
where $d^t = -(\delta I_l + \partial^2 J(\mu^t))^{-1} \bigtriangledown J(\mu^t)$
Until $t \geq$ max_iter or $\| \mu^t - \mu^{t+1} \|_2 \leq$ tol
$w = \frac{1}{\epsilon}(A^T D\mu - 1_n)_+$

Following the definition of generalized Hessian in [74], the gradient and hessian for Eqn. 37 are given as,

$$\begin{aligned} \bigtriangledown J(\mu) &= -\epsilon 1_l + DA(A^T D\mu - 1_n)_+ + (\mu - \pi 1_l)_+ - (-\mu)_+ \\ \partial^2 J(\mu) &= DA\mathrm{diag}\{(\mathrm{A^T D}\mu - 1_\mathrm{n})_*\}\mathrm{A^T D} + \mathrm{diag}\{(\mu - \pi 1_\mathrm{l})_* + (-\mu)_*\} \end{aligned}$$

where $u_* = \mathrm{Sign}(\mathrm{u}_+)$, with Sign being applied element-wise on the vector.

Notice that at each Newton iteration, we need to invert the matrix $Q = \delta I_l + \partial^2 J(\mu)$. This is computationally expensive when the total number of data points l is large ($l > 1000$). We address this issue by using the Sherman-Morrison-Woodbury formula [39] and reduce the time complexity from $O(l^3)$ to $O(ln^2) + O(n^3)$. The inversion of the hessian matrix Q can be computed as follows,

$$\begin{aligned} Q &= F + H * H^T \\ Q^{-1} &= F^{-1} - F^{-1}H(I_l + H^T F^{-1} H)^{-1} H^T F^{-1} \end{aligned}$$

where diagonal matrix $F = \mathrm{diag}(\rho)$ with $\rho = \delta 1_l + (\mu - \nu 1_l)_* + (-\mu)_*$ and $\rho > 0$, and matrix $H = DAE$ with $E = (\mathrm{diag}(\mathrm{A^T D}\mu - 1_\mathrm{n})_*)^{\frac{1}{2}}$.

4.4 Results and Discussion

4.4.1 Data Set Description

The dataset used in this study consists of the values of various energy terms for a non-redundant set of 171 proteins that fall into the *ab initio* folding class. This set is representative of the "hard target" protein sequences in the Protein Data Bank with up to 200 residues, meaning that current homology search tools fail to identify proteins with an evolutionary relationship with proteins in this class.

For each protein, a large set of non-native random conformations (around $51,000$ to $63,000$ per protein) are generated in the manner described in [121]. The energy terms for the native structure and each one of the generated conformations are collected. The energy terms are obtained from the CABS (C_α-C_β-Side chain) force

field [121], which is used in the protein structure prediction tool TASSER [123]. We include 20 different energy terms from this force field, briefly summarized in Table 18.

Table 18: Energy Terms used in TASSER

$E_{*,1}$	pairwise interaction of C_α-SC (side chain)
$E_{*,2}$	pairwise interaction for non-parallel C_α-C_α
$E_{*,3}$	excluded volume of SC-SC
$E_{*,4}$	pairwise interaction of SC-SC
$E_{*,5}$	quarsi3 for SC-SC
$E_{*,6}$	enhance good piece
$E_{*,7}$	-1/r for parallel contact of C_α-C_α
$E_{*,8}$	hydrogen bond interactions on the alpha helix
$E_{*,9}$	hydrogen bond interactions on the beta sheet
$E_{*,10}$	bury potential for SG (side group)
$E_{*,11}$	bias2,3 : $v(i) - v(i+4)$ anti/parallel $c(i) - c(i+2)$ anit/paralel
$E_{*,12}$	crumpling
$E_{*,13}$	bias4 to predicted secondary structure
$E_{*,14}$	bias1 to possible secondary structure
$E_{*,15}$	correlation of E13 of C_α
$E_{*,16}$	correlation of E14
$E_{*,17}$	correlation of E15
$E_{*,18}$	environment potential
$E_{*,19}$	deviation from predicted contact order
$E_{*,20}$	deviation from predicted contact number

The structural similarity of conformations is measured by the TM-score [122], which is intended as a more accurate similarity measure than the commonly used root-mean-squared distance (RMSD) [62] between the the conformations. The values of the TM-score range between 0 and 1, with 1 corresponding to a perfect match between two conformations. We are seeking a large correlation between structural *dis*similarity from the native structure, so we work with 1-(TM-score) instead of the TM-score itself.

4.4.2 Previous Approach

In an earlier optimization study [121], the authors proposed to use an objective function ($G1 * G3$) related to the correlation corr($r(q)$, $E(q)$) between the structural dissimilarity and the total energy of the generated conformations. Namely, they used the product of two quantities $G1$ and $G3$, given by,

$$\begin{aligned} G1 &= \frac{1}{1+\frac{1}{N}\sum_{q=1}^{N} \mathrm{corr(r(q),E(q))}} \\ G3 &= \frac{1}{1+\frac{1}{N}\sum_{q=1}^{N} Z_n(E(q))} \end{aligned}$$

where $Z_n(E(q)) = (\bar{E}(q) - E_0(q))/(\sqrt{\bar{E^2}(q) - (\bar{E}(q))^2})$ is Z-score of the mean of the total energy.

Using the CERN MINUIT package [58] to optimize the weights, they achieved significant results in CASP [123],[125]. Their study employed proteins from all homology modeling, threading, and *ab initio* prediction classes.

4.4.3 Experiment Design

The number of all pairwise difference vectors $z_{ij} = x_i - x_j$ is quadratic in the number of data points (conformations). In addition to this computational issue, it is not realistic to expect the energy function to rank all conformations according to their dissimilarity from the native structure. Thus, instead of working with all possible pairs of conformations, we design a sampling approach as follows:

- For the first class, we sample 100 non-native conformations and include their comparisons with the native structure, i.e., $C_1 = \{z_{i0} = x_i - x_0 \mid y_{i0} = \mathrm{sign(r_i - r_0)} = 1\}$. We sort the set of conformations by dissimilarities to the native structure and perform a uniform sampling.

- For the second class, we generate pairs of comparisons between non-native structures. If two conformations have close values of dissimilarity from the native structure, it may not be reasonable to require the energy function to rank them

according to the dissimilarity. After all, conformations with very different structures can have close values of dissimilarity to the native structure, and in such a case it is not easy to clearly identify which one is "better". For this reason, we restrict the second class to pairs whose dissimilarities from the native structure are sufficiently different. In particular, we first partition the set of non-native conformations into 6 subsets, $S_{(0,0.1)}$, $S_{[0.1,0.2)}$, $\cdots$, $S_{[0.4,0.5)}$, $S_{[0.5,0.6]}$, where $S_{(0,0.1)}$ contains conformations with dissimilarity from the native structure in the range $(0, 0.1)$, $S_{[0.1,0.2)}$ contains conformations with dissimilarity in the range$[0.1, 0.2)$, etc. We then uniformly sample 25 conformations $\{s_i^{(j)}\}_{i=1}^{25}$ from each subset $S_{[a_j,b_j)}$, and sort them according to dissimilarity. The comparisons we include are then, $(s_1^{(1)} - s_1^{(3)}), (s_2^{(1)} - s_2^{(3)}), \ldots, (s_{25}^{(1)} - s_{25}^{(3)}), (s_1^{(2)} - s_1^{(4)}), (s_2^{(2)} - s_2^{(4)}), \ldots, (s_{25}^{(2)} - s_{25}^{(4)}), \ldots, (s_{25}^{(4)} - s_{25}^{(6)})$. For each of these comparisons, the class $y_{ij} = \mathrm{sign}(\mathrm{r_i} - \mathrm{r_j}) = -1$. By picking the pairs in this way, we make sure that the minimum difference between the dissimilarities of any pair included is at least 0.1.

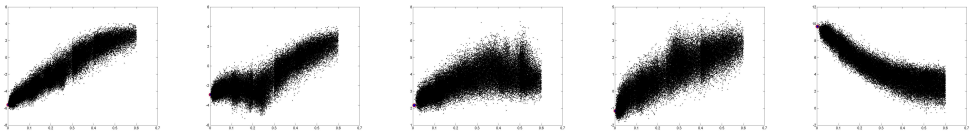
One may ask whether the class labels ± 1 matter; by reversing the order of the conformations in any given pair, the corresponding data point for the SVM changes its class. How should one pick which order to use? It turns out that this order/class assignment is irrelevant, since for the ranking SVM which has no offset, one ends up with exactly the same optimization problem after such a reversal.

By the sampling method described above, we generate 100 data points in each class, for each protein. This gives a total of $34,200$ data points of dimension 20.

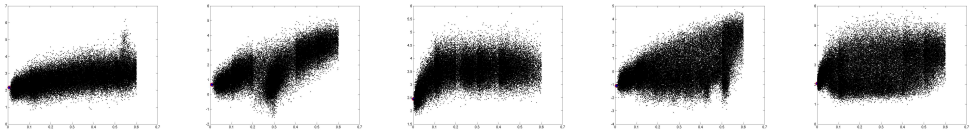
4.4.4 Performance Measure

In order to compare alternative methods of energy function optimization, we need appropriate criteria. The ultimate test is, of course, folding; given two energy functions, one runs a folding algorithm with each in order to see which one performs better for

a test set of proteins. We leave this ultimate test to our future work, and describe criteria that are suggestive of the expected performance of energy functions/weights.



(a) 1tluA (b) 1ayoA (c) 1vk5A (d) 1kafA (e) 1dwnA



(f) 1pp7U (g) 1xzpB (h) 1b79A (i) 1azzC (j) 1d7mA

Figure 14: Representative Energy versus Structural Similarity (1-(TM-score)) Plots

As we have suggested in the introduction, the performance of a protein energy function can be evaluated descriptively through plots of the total energy versus structural dissimilarity for each protein of the test set in a given cross-validation fold. The plots can illustrate possible *ab initio* folding paths of a protein, starting from a random conformation with high structural dissimilarity, moving to the native structure or near-native conformations by reducing the value of the energy function. In order to give some visual indication of the quality of the energy function resulting from our optimization, we display in Figure 14 ten representative test-set plots of total energy versus dissimilarity (1-TM score) for weights obtained using RankingSVM^{NNL_1}. Each dot represents a decoy structure, and the thick red point (roughly in the bottommost, leftmost end of the plot) represents the native structure.

Desirable results are shown in (a) and (d) for proteins 1tluA and 1kafA, respectively. Due to the consistent, monotonic trend, reducing the total energy by changing the conformation corresponds to increasing the similarity to the native structure. Thus, one expects the learned energy function to result in rapid *ab inito* folding from a random conformation. Plots (b) 1ayoA and (g) 1xzpB show good, monotonic regions together with highly undesirable, deep local minima where the folding procedure

may get stuck. Plots (c) 1vk5kA, (f) 1abc_ and (h) 1b79A show good behavior near the native structure, leading an energy minimization procedure towards the native structure, while having trouble for random structures that are highly dissimilar to the native structure. (e) 1dwnA and (j) 1d7mA represent highly undesirable cases, where there is no observable relation and negative correlation between the dissimilarity and energy, respectively.

While such plots are descriptive of the general characteristics of an energy function, one would like to have a quantitative measure of performance. In the previous approach used in TASSER [123], the objective function $G1 * G3$ we described above was used as the performance measure. However, this is an ad hoc objective function whose direct relationship with folding performance is not clear.

In the following, we use two criteria to evaluate the fitness of the learned energy functions. The first one, Pearson's correlation, measures the strength of the linear relationship between the energy and the structural dissimilarity. We expect the relationship between energy and dissimilarity to be nonlinear, with outliers, unequal variances, and non-normality. Thus, while it is a useful first test of an overall linear dependence, we do not expect Pearson's correlation to give an accurate prediction of the folding performance that will be achieved by using a given an energy function.

The second approach we use for evaluating energy functions is more flexible, and possibly more useful: we compare the rankings of the conformations provided by the total energy and the dissimilarity from the native structure. Since ab-initio folding approaches work by modifying conformations to reduce their total energy, we can expect to obtain accurate folds when reducing the total energy corresponds to reducing the dissimilarity to the native structure, i.e., when the relative order of two conformations given by the energy and the dissimilarity agree. This approach does not give a particular importance to linear relations; all that matters is the degree to which there is a monotonic relation between the energy and the dissimilarity to the

native structure.

One of the most popular statistics used to compare two rankings of a set is Kendall's τ statistic [64]. This defines the distance between two rankings of a set as the number of pairs that have the opposite ordering in the two rankings. In our problem, the two rankings of the collected conformations are given by the total energy and the dissimilarity from the native structure. We approximate the expected value of Kendall's τ statistic for the these two ranking functions through the sampled pairwise agreement between the ranks they assign.

4.4.5 Results Analysis

We evaluate our RankingSVM approach to learning protein energy functions through 10-fold cross validation. We randomly partition the 171 proteins in our data set into 10 folds. For each fold i, we learn an energy function from the energy data of the other 9 folds, and evaluate the learned energy function on fold i. As described above, we use two normalization schemes in the optimization of the NNL2SVM method. We denote ranking via NNL2SVM with the normalization rule $||\ w^t\ ||_1 = 1$ as NNL2SVM$_{n1}$, and ranking via NNL2SVM with the normalization rule $||\ w^{t+1}\ ||_1 \leq ||\ w^t\ ||_1$ as NNL2SVM$_{n2}$. Ranking via the NNL1SVM method is denoted as NNL1SVM.

The baseline method in our experiments consists of optimizing the objective function $G1 * G3$ using the MINUIT package. Since $G1 * G3$ is not a convex function, the solution depends heavily on the initial weights. TASSER protein structure prediction program sets the initial weights based on domain knowledge on the importance of the energy terms, as well as the proteins in the training set. This initialization procedure will be too expensive for our random 10-fold cross validation. Therefore, in our experiments, we tried 1000 sets of initial weights randomly generated from the set $\{0, 1\}^{20}$, and then selected the ones with the smallest $G1 * G3$ value (we remove the results that having $Z_n(w) \approx -1$).

4.4.5.1 Parameters Tuning

We employ the grid search procedure for choosing the right parameter setting for NNL2SVM and NNL1SVM methods. Note that for NNL2SVM method, we need to tune the parameter $\nu > 0$, and for NNL1SVM method, the tuning parameter are $\pi > 0$ and $\delta > 0$. The candidate values of the parameters used for the experiments were

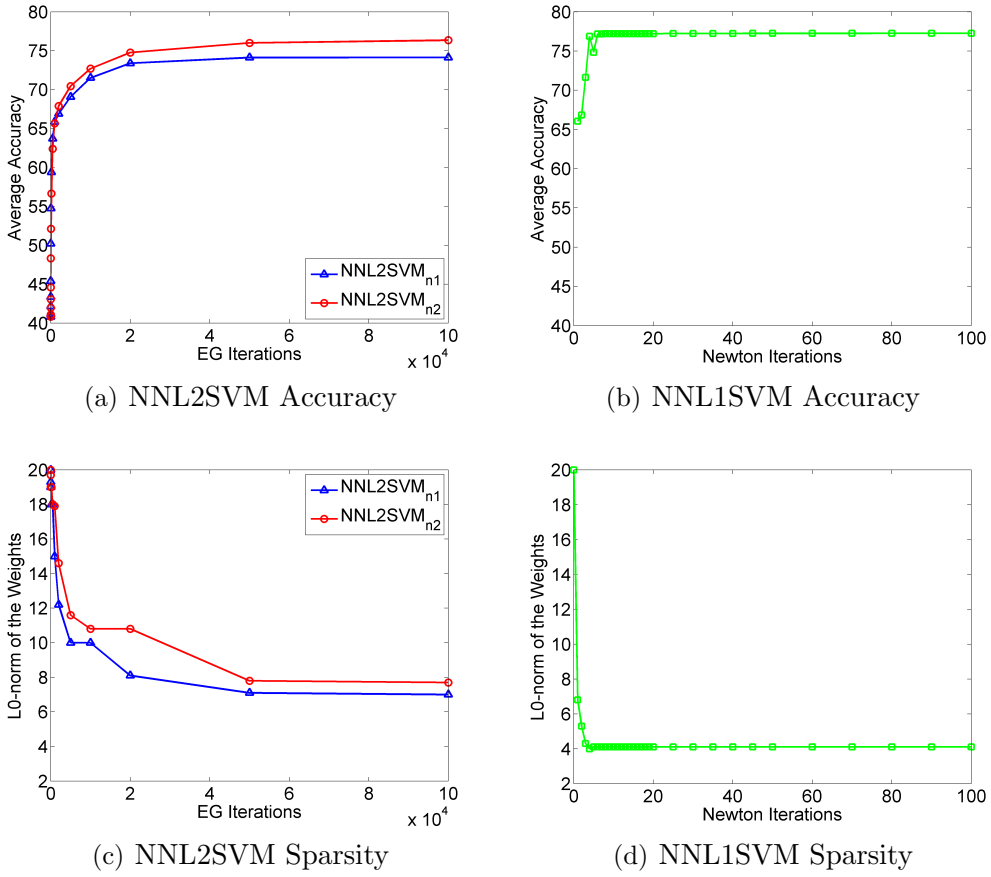
- $\nu, \pi \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$
- $\delta \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$

For each parameter setting, we perform a 10-fold cross-validation and obtain the average score over the 10 folds, the averaged training performance over cross-validation. The parameter setting with the best score (ties are broken randomly) is chosen for the final comparison.

4.4.5.2 NNL2SVM versus NNL1SVM

We first analyze the trend of the sampled pairwise agreement and the sparsity of the weight vector during the algorithm optimization of the proposed NNSVMs. For NNL2SVM, we observe the testing accuracy and the sparsity at iterations {1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000, 100000} during the exponential gradient (EG) optimization. For NNL1SVM, we observe at iterations {1, $\cdots$, 20, 25, 30, 35, 40, $\cdots$, 100} during the Newton optimization.

As shown in Figure 15(a), NNL2SVM$_{n_2}$ (shown as red squares) outperforms NNL2SVM$_{n_1}$ (blue triangles). While the latter converges earlier, it does so at the cost of lower accuracy. NNL1SVM (Figure 15(b)) converges only after 20 Newton iterations. Unlike the NNL2SVM methods, its performance during optimization is irregular. But it still indicates an overall trend of increasing accuracy.



(a) NNL2SVM Accuracy (b) NNL1SVM Accuracy

(c) NNL2SVM Sparsity (d) NNL1SVM Sparsity

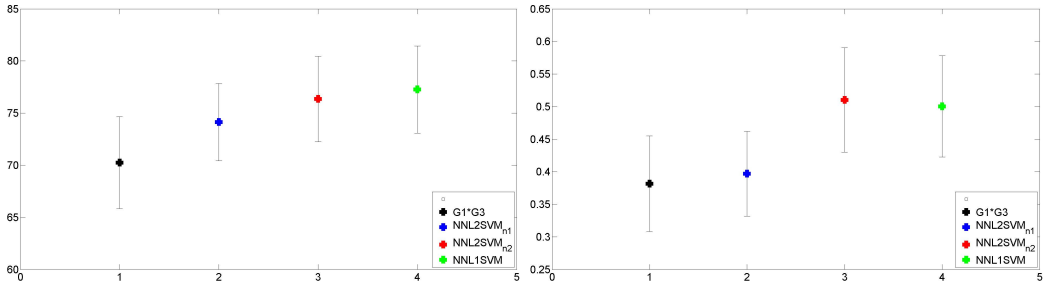
Figure 15: NNSVM Optimization Methods Comparison

In Figure 15(c), (d), we see that NNL1SVM achieve low sparsity after only a few Newton iterations, while NNL2SVM methods gradually obtain sparse solutions after about 50000 EG iterations. NNL2SVM$_{n_1}$ generally obtains sparser solutions than those of NNL2SVM$_{n_2}$, but at the cost of accuracy. The nonzero energy terms chosen are as follows. For NNL2SVM$_{n_1}$, the consensus is terms 2, 7, 9, 12, 18, 19, 20. The results from NNL2SVM$_{n_2}$ agree, with addition of energy term 16. For NNL1SVM, the preferred terms are 2, 9, 18, 20. The biochemical meaning of these combinations are under investigation.

4.4.5.3 RankingSVM versus TASSER$^{\text{MINUIT}}$

We then measures the performance of the energy function learned by each method using both Kendall τ statistics (approximated by sampled pairwise agreement) and

Pearson's correlation. Figure 16(a) lists the sampled pairwise agreement, measured by testing accuracy on the labeled pairwise difference data over the 10-fold cross validation. Figure 16(b) list the average correlation coefficients between the rmsd value and the energy, which are computed using the learned energy function during each cross-validation.



(a) Pairwise Agreement (b) Correlation

Figure 16: Error Plot of the Performance of the Learned Energy Functions

Comparing to the energy function learned by baseline method, the energy functions learned using RankingSVM methods generally have better performance in both sampled pairwise agreement and correlation. The energy functions learned using RankingSVM$^{\text{NNL2}}$ method can achieve around 8.7% increase in the sampled pairwise agreement and around 33.7% increase on the correlation values on average, while those output by RankingSVM$^{\text{NNL1}}$ method have around 9.7% and 28.3% increase on those values, respectively. In addtion, the average computation time of a 10-fold cross validation for RankingSVM$^{\text{NNL1}}$ is about 7 seconds, which is much more efficient comparing to RankingSVM$^{\text{NNL2}}$ methods (around 20 minutes) and the baseline method (around 2 minutes). Overall, we can conclude that RankingSVM$^{\text{NNL1}}$ method outperforms the other methods in terms of learning performance as well as efficiency.

4.4.6 Discussion

A critical open problem in *ab initio* protein folding is protein energy function design. We address this problem as a weight optimization problem, and demonstrate a

machine learning approach using the ranking-via-classification paradigm. Due to the nature of the commonly used data generation procedures, the sampling of the non-native conformations used in the training set is restricted to a small subspace of the space of all possible conformations consisting of compact configurations. In order to avoid a tendency to overfit to this subspace and preserve the physicality of the energy function, we restrict the problem to non-negativity weights. We develop two efficient algorithms, NNL2SVM and NNL1SVM, to solve the constrained RankingSVM problem. Comparing with state-of-the-art approach that based on maximizing the correlation between the total energy and the structural dissimilarity, our learning-to-rank approach was able to learn energy functions that maintain the correct ordering of the conformations more often, and give higher correlations with the dissimilarity from the native structure. In addition, NNL1SVM, the method with the highest accuracy, is much more efficient for learning on a large protein set.

CHAPTER V

CONCLUSION

Bioinformatics is a growing application area of machine learning. Many of the computational biology problems can be reduced into supervised learning problems. In this thesis, we extend support vector machine optimization, which has been show superior classification performance theoretically and empirically, to develop efficient algorithms on the reduced machine learning problems and provide useful tool to assist the biologists. To examplify our approach, we solve problems from three classes of important bioinformatic application including cancer diagnosis, biomarker discovery and protein energy function learning.

We investigate predictiveness of metabolic profiles generated by Direct Analysis in Real Time (DART) from patient sera. We reduce the task of classifying DART MS profiles into a functional classification problem and solve it using functional SVM method. The assay distinguished between the cancer and control groups with 98.9% accuracy (100% sensitivity; 98-100% specificity) under leave-one-out cross-validation evaluation. We view this as a successful step towards the development of an accurate new approach to the diagnosis of ovarian and other cancers.

We explore convex relaxations of L_0-norm SVM for developing more aggressive feature selection methods to assist high throughput biomarker discovery. We study mixed-integer support vector machine and solve it with mixed-integer nonlinear programming technique. Empirical study shows encouraging result, simultaneously improved feature selection and prediction performance on small/medium feature size data sets. We also propose an practical solution to solve the fractional-norm SVM problem with difference of convex programming technique. Empirical study support

the effectiveness of the fractional-norm SVM over other commonly-used sparse SVM methods. We believe our approach is a promising direction for feature selection in high-dimensional low-sample size data sets

We address the open problem of learning energy function for *ab initio* protein folding as a weight optimization problem and demonstrate a machine learning approach, learning to rank, to solve this problem. To maintain the physicality of the results, we impose non-negativity constraints on the weights and develop two non-negative support vector machine methods (NNL2SVM and NNL1SVM) for the constrained RankingSVM optimization. Comparing with the state-of-the-art approach, our methods result in energy functions that maintain the correct ordering of the conformations more often, and give higher correlations with the dissimilarity from the native structure. Furthermore, the ability to learn SVMs with non-negative weights is a more general capability which we anticipate have applications beyond protein folding.

5.1 Future Work

Incorporate Domain Knowledge into Biomarker Discovery

Feature selection methods that utilize cancer data (e.g. gene expression data, or metabolite intensity data) alone may be insufficient to produce biologically compelling biomarker candidates. The main reason is that these data sets only monitor one level of biological regulations, of which there are many levels in cancer pathogenesis. Furthermore, many of these low sample size high dimensional data sets are linearly separable between different patient groups, hence the classification problem on these datasets is too simple, admitting too many possible solutions in such a high-dimensional space for us to be able to pinpoint critical features. Changes on the training data subsampling often results in completely different feature selection results. One possible solution to boost the biological significance and stability of feature selection results would be incorporating domain knowledge, which is generally used

in the biological validation process, into the feature selection process.

We have explored on incorporating gene ontology information [5, 29] into the statistical microarray analysis [42]. The method can be described as in Table 19. We conduct experiments on a $43 \times 12,558$ ovarian cancer microarray data generated from Affymetrix U95Av2 chips (containing 10 are benign cancer patients; 9 are malignant cancer patients with no chemotherapy treatment; 24 are malignant cancer patients with chemotherapy treatment). The results showed that this method is capable of recovering biomarkers such as TUMOR PROTEIN 53 (TP53), a verified biomarker for ovarian cancer, whose expression values are not significantly different between patient groups, but instead may be mutated or regulated at the post-translational level through ontological links from geno ontology.

Table 19: Incorporating Gene Ontology into Biomarker Discovery

Divide genes into *function groups* according to gene annotations
Compute the discriminative capability of each *function group*
Obtain the gene expression submatrix of the jth *function group*
Compute SVM accuracy on the gene expression submatrix
Compute SVM accuracy during feature selection process on the data set
Score the *function group* with $LOOCV\_full$ and $best\_LOOCV$
Rank genes according to the scores of the functional group it belongs to
Normalize on the gene score (Optional)

Still, further investigations are required in order to apply this idea to biomarker discovery in general. For example for metabolites panel selection, many metabolite might not be annotated since metabolite cancer biomarker is not as extensive studied as large polymers cancer biomarkers (such as genes). In addition, the annotation on the mass spectrum is not as straightforward as gene annotation and still requires a lots of input from domain experts, for example, mapping m/z values into mebabolite mass for metabolite identification. We would also like to explore other ontological information (such as KEEG [63]) to see their effect in boosting the biological significance and stability of feature selection results.

Energy Function Learning

While our proposed non-negative RankingSVM approach is a promising direction for learning energy function for *ab initio* protein folding, further improvements might be needed to assist the biological validation task: folding proteins. As the rankingSVM problem is contrained with non-negativity weights, many of the weight elements are threshold to zero during the optimization. As the data set only consists of a small subset of all the possible conformations of a given protein, energy function with just a few energy terms (i.e., sparse weight vector) might have a tendency to overfit to this subspace. It is possible that the energy function performs well on the chosen set of existing proteins, while perform poorly when predicting new physical structures. Thus, denser solution of the weight vector is preferred in order to avoid overfitting to the sampled conformations.

We can address this issue by constraining the rankingSVM problem with a more general boundary condition $w > w_0$, where the boundary weight values $w_0 \neq 0$ can be determined using domain knowledge on the energy terms. Given the updated optimization formulation described in Table 20 and the boundary vector w_0, we can easily modify our optimization algorithms (see Table 16,17) to solve the revised rankingSVM problems with general boundary conditions.

Table 20: Ranking SVM with General Boundary Constraints

min $\frac{\nu}{2} w^T w + \frac{1}{2n} \mid\mid (1_n - DAw)_+ \mid\mid_2$ s.t. $w \geq w_0$	min $1_d^T w + c 1_n^T (1_n - DAw)_+$ s.t. $w \geq w_0$

REFERENCES

[1] Ahmed, N., Oliva, K., Barker, G., Hoffmann, P., Reeve, S., Smith, I., Quinn, M., and Rice, G., "Proteomic tracking of serum protein isoforms as screening biomarkers of ovarian cancer," *Proteomics*, vol. 5, no. 17, 2005.

[2] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the national academy of sciences of the United States of America*, vol. 96, no. 12, p. 6745, 1999.

[3] Amaldi, E. and Kann, V., "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 1-2, pp. 237–260, 1998.

[4] Anderson, N. and Anderson, N., "The human plasma proteome: history, character, and diagnostic prospects.," *Molecular & cellular proteomics: MCP*, vol. 1, no. 11, p. 845, 2002.

[5] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Traver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., and Rubin, G., "Gene ontology: tool for the unification of biology gene ontology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[6] Atkinson, A., Colburn, W., DeGruttola, V., DeMets, D., Downing, G., Hoth, D., Oates, J., Peck, C., Schooley, R., Spilker, B., and others, "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework*," *Clinical Pharmacology & Therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[7] Baker, D., Morrison, P., Miller, B., Riely, C., Tolley, B., Westermann, A., Bonfrer, J., Bais, E., Moolenaar, W., and Tigyi, G., "Plasma lysophosphatidic acid concentration and ovarian cancer," *J Am Med Assoc*, vol. 287, no. 23, pp. 3081–3082, 2002.

[8] Barker, M. and Rayens, W., "Partial least squares for discrimination," *J Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[9] Ben-Israel, A. and Greville, T., *Generalized inverses: Theory and applications*. Springer Verlag, 2003.

[10] Biau, G., Bunea, F., and Wegkamp, M., "Functional classification in Hilbert spaces," *Information Theory, IEEE Transactions on*, vol. 51, no. 6, pp. 2163–2172, 2005.

[11] Bonami, P., Biegler, L., Conn, A., Cornuéjols, G., Grossmann, I., Laird, C., Lee, J., Lodi, A., Margot, F., Sawaya, N., and others, "An algorithmic framework for convex mixed integer nonlinear programs," *Discrete Optimization*, vol. 5, no. 2, pp. 186–204, 2008.

[12] Bowen, N., Walker, L., Matyunina, L., Logani, S., Totten, K., Benigno, B., and McDonald, J., "Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells," *BMC medical genomics*, vol. 2, no. 1, p. 71, 2009.

[13] Bradley, P. and Mangasarian, O., "Feature selection via concave minimization and support vector machines," in *Machine Learning Proceedings of the Fifteenth International Conference (ICML98)*, pp. 82–90, 1998.

[14] Braga-Neto, U. and Dougherty, E., "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics*, vol. 20, no. 3, p. 374, 2004.

[15] Breiman, L., "Heuristics of instability and stabilization in model selection," *The Annals of Statistics*, vol. 24, no. 6, pp. 2350–2383, 1996.

[16] Brown, M. L., Riley, G. F., Schussler, N., and Etzioni, R. D., "Estimated health care costs related to cancer treatment from SEER-Medicare data," *Med Care*, vol. 40(8 Supplement)IV, pp. 104–117, 2002.

[17] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G., "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*, p. 96, 2005.

[18] Butterworth, A., "Family history as a risk factor for common, complex disease. An independent, epidemiologic assessment of the evidence for familial risk disease. Cambridge, UK," *Public Health Genetics Foundation*, 2007.

[19] Candes, E., Wakin, M., and Boyd, S., "Enhancing sparsity by reweighted L1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[20] Cao, Z., Qin, T., Liu, T., Tsai, M., and Li, H., "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*, p. 136, 2007.

[21] Caruana, R., Baluja, S., and Mitchell, T., "Using the future to sort out the present: Rankprop and multitask learning for medical risk analysis," *Neural Information Processing 7*, 1995.

[22] Cascino, A., Muscaritoli, M., Cangiano, C., Conversano, L., Laviano, A., Ariemma, S., Meguid, M., and Fanelli, F., "Plasma amino acid imbalance in patients with lung and breast cancer," *Anticancer research*, vol. 15, no. 2, pp. 507–510, 1995.

[23] Centers for Disease Control, United States Cancer Statistics . http://apps.nccd.cdc.gov/uscs/.

[24] Chang, C.-C. and Lin, C.-J., *LIBSVM: a library for support vector machines*, 2001. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[25] Chang, K., Hsieh, C., and Lin, C., "Coordinate descent method for large-scale l2-loss linear support vector machines," *J. Machine Learning Res.*, vol. 9, pp. 1369–1398, 2008.

[26] Cody, R., "Observation of molecular ions and analysis of nonpolar compounds with the direct analysis in real time ion source," *Analytical Chemistry*, vol. 81, no. 3, pp. 1101–1107, 2008.

[27] Cody, R., Laramée, J., and Durst, H., "Versatile new ion source for the analysis of materials in open air under ambient conditions," *Analytical Chemistry*, vol. 77, no. 8, pp. 2297–2302, 2005.

[28] Conrads, T., Fusaro, V., Ross, S., Johann, D., Rajapakse, V., Hitt, B., Steinberg, S., Kohn, E., Fishman, D., Whitely, G., and others, "High-resolution serum proteomic features for ovarian cancer detection," *Endocrine-Related Cancer*, vol. 11, no. 2, pp. 163–178, 2004.

[29] Consortium, G. O., "The gene ontology (go) database and informatics source," *Nucleic Acids Research*, vol. 32, no. Database Issue, 2004.

[30] Cossock, D. and Zhang, T., "Subset ranking using regression," *Lecture Notes in Computer Science*, vol. 4005, p. 605, 2006.

[31] Crammer, K. and Singer, Y., "Pranking with ranking," *Advances in neural information processing systems*, vol. 1, pp. 641–648, 2002.

[32] Denkert, C., Budczies, J., Kind, T., Weichert, W., Tablack, P., Sehouli, J., Niesporek, S., Könsgen, D., Dietel, M., and Fiehn, O., "Mass Spectrometry–Based Metabolic Profiling Reveals Different Metabolite Patterns in Invasive Ovarian Carcinomas and Ovarian Borderline Tumors," *Cancer research*, vol. 66, no. 22, p. 10795, 2006.

[33] Dettmer, K., Aronov, P., and Hammock, B., "Mass spectrometry-based metabolomics," *Mass Spectrom Rev*, vol. 26, no. 1, pp. 51–78, 2007.

[34] Fan, J. and Li, R., "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties.," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1361, 2001.

[35] Freund, Y., Iyer, R. D., Schapire, R. E., and Singer, Y., "An efficient boosting algorithm for combining preferences," *Machine Learning Research*, vol. 4, pp. 933–969, 2003.

[36] Friesner, R. and John, R., "Computational studies of protein folding," *Annual Review of Biophysics & Biomolecular Structure*, vol. 25, pp. 315–342, 1996.

[37] Fung, G. and Mangasarian, O., "A feature selection newton method for support vector machine classification," *Computational optimization and applications*, vol. 28, no. 2, pp. 185–202, 2004.

[38] Gasso, G., Rakotomamonjy, A., and Canu, S., "Recovering sparse signals with a certain family of non-convex penalties and DC programming," *IEEE Trans. Signal Processing, To appear*, 2009.

[39] Golub, G. and Van Loan, C., *Matrix computations*. Johns Hopkins Univ Pr, 1996.

[40] Guan, W., Arkadas, O., Gray, A., Borreguero, J., Pandit, S., Jagielska, A., Wroblewska, L., and Skolnick, J., "Learning Protein Folding Function," 2011. http://www.cc.gatch.edu/∼wguan/learning_protein_folding_function.pdf.

[41] Guan, W. and Gray, A., "Sparse High-dimensional Fractional-norm Support Vector Machine Via DC Programming," *Computational Statistics and Data Analysis*, 2011.

[42] Guan, W., Gray, A., Navathe, S., Bowen, N., and McDonald, J., "Discovering Ovarian Cancer Biomarkers using Gene Ontology based Microarray Analysis," *Workshop on Data Mining for Bioinformatics*, 2007.

[43] Guan, W., Gray, A., and S., L., "Mixed Integer Support Vector Machine," *NIPS workshop on Optimization for Machine Learning*, 2009.

[44] Guan, W., Zhou, M., Hampton, C., Benigno, B., Walker, L., Gray, A., McDonald, J., and Fernández, F., "Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines," *BMC bioinformatics*, vol. 10, no. 1, p. 259, 2009.

[45] Günlük, O. and Linderoth, J., "Perspective relaxation of mixed integer nonlinear programs with indicator variables," in *Proceedings of the 13th international conference on Integer programming and combinatorial optimization*, pp. 1–16, Springer-Verlag, 2008.

[46] Guyon, I. and Elisseeff, A., "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[47] Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G., "Result analysis of the nips 2003 feature selection challenge," *Advances in Neural Information Processing Systems*, vol. 17, pp. 545–552, 2005.

[48] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[49] Harris, G., Nyadong, L., and Fernandez, F., "Recent developments in ambient ionization techniques for analytical mass spectrometry," *Analyst*, vol. 133, no. 10, pp. 1297–1301, 2008.

[50] Herbrich, R., Obermayer, K., and Graepel, T., "Large margin rank boundaries for ordinal regression," in *Advances in Large Margin Classifiers*, pp. 115–132, 2000.

[51] Hingorani, S., Petricoin III, E., Maitra, A., Rajapakse, V., King, C., Jacobetz, M., Ross, S., Conrads, T., Veenstra, T., Hitt, B., and others, "Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse," *Cancer Cell*, vol. 4, no. 6, pp. 437–450, 2003.

[52] Hiriart-Urruty, J. and Lemaréchal, C., *Convex Analysis and Minimization Algorithms: Fundamentals*. Springer, 1993.

[53] HMDB: Human Metabolome Database. http://www.hmdb.ca.

[54] Horner, M., Ries, L., Krapcho, M., Neyman, N., Aminou, R., Howlader, N., Altekruse, S., Feuer, E., Huang, L., Mariotto, A., and others, "SEER cancer statistics review, 1975-2006," *Bethesda, MD: National Cancer Institute*, p. 34, 2009. http://seer.cancer.gov/csr/1975_2006/.

[55] IBM ILOG CPLEX package. http://www-01.ibm.com/software/integration/optimization/cplex-optimizer.

[56] Iyer, R., Lewis, D., Schapire, R., Singer, Y., and Singhal, A., "Boosting for document routing," in *Proceedings of the 19th CIKM conference*, pp. 70–77, 2000.

[57] Jacobs, I. and Menon, U., "Progress and challenges in screening for early detection of ovarian cancer," *Molecular & Cellular Proteomics*, vol. 3, no. 4, p. 355, 2004.

[58] James, F., Winkler, M., and CERN, G., "MINUIT Package, CERN, Geneva," 2004. http://seal.web.cern.ch/seal/snapshot/work-packages/ mathlibs/minuit/.

[59] Joachims, T., "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, ACM, 2002.

[60] Jobson, J., *Applied multivariate data analysis*. Springer, 1991.

[61] Johnson, R. and Wichern, D., *Applied multivariate statistical analysis*, vol. 5. Prentice Hall Upper Saddle River, NJ, 2002.

[62] Kabsch, W., "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 32(5), pp. 922–923, 1976.

[63] KEGG: Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp.

[64] Kendall, M., "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.

[65] Kivinen, J. and Warmuth, M., "Exponentiated gradient versus gradient descent for linear predictors," *Information and Computation*, 1997.

[66] Knight, K. and Fu, W., "Asymptotics for lasso-type estimators," *Annals of Statistics*, pp. 1356–1378, 2000.

[67] Le Thi, H., Le, H., Nguyen, V., and Pham Dinh, T., "A DC programming approach for feature selection in support vector machines learning," *Advances in Data Analysis and Classification*, vol. 2, no. 3, pp. 259–278, 2008.

[68] Le Thi, H. and Pham Dinh, T., "A continuous approach for the concave cost supply problem via DC programming and DCA," *Discrete Applied Mathematics*, vol. 156, no. 3, pp. 325–338, 2008.

[69] Lee, J., Chen, M., Chang, C., Tiai, Y., Lin, P., Lai, H., and Wang, S., "Plasma amino acid levels in patients with colorectal cancers and liver cirrhosis with hepatocellular carcinoma," *Hepato-gastroenterology*, vol. 50, no. 53, pp. 1269–1273, 2003.

[70] Leyffer, S., "User manual for MINLP BB," *University of Dundee Numerical Analysis Report*, vol. 234, 1999.

[71] Li, J., Zhang, Z., Rosenzweig, J., Wang, Y., and Chan, D., "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer," *Clinic Chem*, vol. 48, no. 8, pp. 1296–1304, 2002.

[72] Lin, D., Pitler, E., Foster, D., and Ungar, L., "In defense of l0," in *Workhsop on Feature Selection at International Conference on Machine Learning, ICML*, Citeseer, 2008.

[73] Liu, Y., Zhang, H., and others, "Support vector machines with adaptive Lq penalty," *Computational Statistics & Data Analysis*, vol. 51, no. 12, pp. 6380–6394, 2007.

[74] Mangasarian, O., "Exact 1-norm support vector machines via unconstrained convex differentiable minimization," *J Machine Learning Res*, vol. 7, no. 2, pp. 1517–1530, 2006.

[75] Medina, M., Quesada, A., Núñez de Castro, I., and Sánchez-Jiménez, F., "Histamine, polyamines, and cancer," *Biochemical pharmacology*, vol. 57, no. 12, pp. 1341–1344, 1999.

[76] Morris, D. and Adams, W., "Cimetidine and colorectal cancer–old drug, new use?," *Nature medicine*, vol. 1, no. 12, p. 1243, 1995.

[77] Mueller, W., Handschumacher, R., and Wade, M., "Serum haptoglobin in patients with ovarian malignancies," *Am Coll Obstet Gynecol J*, vol. 38, no. 3, pp. 427–435, 1971.

[78] Murphy, P. and Aha, D., "UCI repository of machine learning databases," 1992.

[79] National Cancer Institute, Cancer Stat Fact Sheets. http://seer.cancer.gov/statfacts/html/ovary.html.

[80] Odunsi, K., Wollman, R., Ambrosone, C., Hutson, A., McCann, S., Tammela, J., Geisler, J., Miller, G., Sellers, T., Cliby, W., and others, "Detection of epithelial ovarian cancer using 1 H-NMR-based metabonomics," *Int J Cancer*, vol. 113, no. 5, pp. 782–788, 2005.

[81] Pan, Z., Gu, H., Talaty, N., Chen, H., Shanaiah, N., Hainline, B., Cooks, R., and Raftery, D., "Principal component analysis of urine metabolites detected by NMR and DESI–MS in patients with inborn errors of metabolism," *Analytical and Bioanalytical Chemistry*, vol. 387, no. 2, pp. 539–549, 2007.

[82] Pearson, H., "Meet the human metabolome," *Nature*, vol. 446, no. 7131, pp. 8–8, 2007.

[83] Petricoin, E., Ardekani, A., Hitt, B., Levine, P., Fusaro, V., Steinberg, S., Mills, G., Simone, C., Fishman, D., Kohn, E., and others, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, no. 9306, pp. 572–577, 2002.

[84] Petru, E., Sevin, B., Averette, H., Koechli, O., Perras, J., and Hilsenbeck, S., "Comparison of three tumor markers–CA-125, lipid-associated sialic acid (LSA), and NB/70K–in monitoring ovarian cancer.," *Gynecol Oncol*, vol. 38, no. 2, pp. 181–186, 1990.

[85] Pham Dinh, T. and Le Thi, H., "A DC optimization algorithm for solving the trust-region subproblem," *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 476–505, 1998.

[86] Rajapakse, J., Duan, K., and Yeo, W., "Proteomic cancer classification with mass spectrometry data," *American Journal of PharmacoGenomics*, vol. 5, no. 5, pp. 281–292, 2005.

[87] Rajaram, S., Garg, A., Zhou, X., and Huang, T., "Classification approach towards ranking and sorting problems," in *Proceedings of the 14th ECML conference*, pp. 301–312, 2003.

[88] Ramsay, J., "Functional Data Analysis Package." Software available at http://www.psych.mcgill.ca/ misc/fda/software.html.

[89] Ramsay, J. and Silverman, B., *Functional Data Analysis*. Springer, 2005.

[90] Rockafellar, R., "Convex Analysis. Princeton landmarks in mathematics," 1997.

[91] Rossi, F. and Villa, N., "Support vector machine for functional data classification," *Neurocomputing*, vol. 69, no. 7-9, pp. 730–742, 2006.

[92] Rui, Z., Jian-Guo, J., Yuan-Peng, T., Hai, P., and Bing-Gen, R., "Use of serological proteomic methods to find biomarkers associated with breast cancer," *Proteomics*, vol. 3, no. 4, 2003.

[93] Saab, R., Chartrand, R., and Yilmaz, O., "Stable sparse approximations via nonconvex optimization," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3885–3888, IEEE, 2008.

[94] Schutter, E., Visser, J., van Kamp, G., Mensdorff-Pouilly, S., van Dijk, W., Hilgers, J., and Kenemans, P., "The utility of lipid-associated sialic acid (LASA or LSA) as a serum marker for malignancy. A review of the literature.," *Tumour Biol: J Int Soc Oncodevelopmental Biol Med*, vol. 13, no. 3, p. 121, 1992.

[95] Schwartz, P., Chambers, S., Chambers, J., Gutmann, J., Katopodis, N., and Foemmel, R., "Circulating tumor markers in the monitoring of gynecologic malignancies," *Cancer*, vol. 60, no. 3, pp. 353–361, 1987.

[96] Schwartz, P. and Taylor, K., "Is early detection of ovarian cancer possible?," *Annals of medicine*, vol. 27, no. 5, pp. 519–528, 1995.

[97] Sieja, K., Stanosz, S., von Mach-Szczypinski, J., Olewniczak, S., and Stanosz, M., "Concentration of histamine in serum and tissues of the primary ductal breast cancers in women," *The Breast*, vol. 14, no. 3, pp. 236–241, 2005.

[98] Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., and others, "Gene expression correlates of clinical prostate cancer behavior," *Cancer cell*, vol. 1, no. 2, pp. 203–209, 2002.

[99] Skolnick, J., Kihara, D., and Zhang, Y., "Development and testing of the PROSPECTOR 3.0 threading algorithm," *Proteins*, vol. 3, pp. 502–518, 2004.

[100] Skolnick, J. and Kolinski, A., "Computational studies of protein folding," *Computing in Science and Enginerring*, vol. 3, pp. 40–48, 2001.

[101] Sreekumar, A., Poisson, L., Rajendiran, T., Khan, A., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R., Li, Y., and others, "Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression," *Nature*, vol. 457, no. 7231, p. 910, 2009.

[102] Sutphen, R., Xu, Y., Wilbanks, G., Fiorica, J., Grendys, E., LaPolla, J., Arango, H., Hoffman, M., Martino, M., Wakeley, K., and others, "Lysophospholipids are potential biomarkers of ovarian cancer," *Cancer Epidem Biomarkers Prevention*, vol. 13, no. 7, pp. 1185–1191, 2004.

[103] Szklo, M. and Nieto, F., *Epidemiology: beyond the basics*. Jones and Bartlett Publishers, 2004.

[104] Tadros, G., Foemmel, R., and Shebes, M., "Plasma lipid-associated sialic acid and serum CA 125 as indicators of disease status with advanced ovarian cancer," *Am Coll Obstet Gynecol J*, vol. 74, no. 3, pp. 379–383, 1989.

[105] Taylor, M., Guiver, J., Robertson, S., and Minka, T., "SoftRank: optimizing non-smooth rank metrics," in *Proceedings of the international conference on Web search and web data mining*, pp. 77–86, 2008.

[106] Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[107] Trygg, J., Holmes, E., and Lundstedt, T., "Chemometrics in metabonomics," *J Proteome Res*, vol. 6, no. 2, pp. 469–479, 2007.

[108] U.S. Preventive Services Task Force. Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: recommendation statement. http://www.ahrq.gov/clinic/uspstf/uspsbrgen.htm.

[109] Vapnik, V., *The nature of statistical learning theory*. Springer, 1995.

[110] Vardi, J., Tadros, G., Malhotra, C., Charney, T., Shebes, M., and Foemmel, R., "Lipid associated sialic acid in plasma in patients with advanced carcinoma of the ovaries.," *Surg Gynecol Obstet*, vol. 168, no. 4, pp. 296–301, 1989.

[111] Villa, N., Rossi, F., and Carcassonne, F., "Recent advances in the use of SVM for functional data classification," in *Proceedings of 1st International Workshop on Functional and Operatorial Statistics (IWFOS 2008), Toulouse, France*, pp. 1–4, 2008.

[112] Wagner, J., Williams, S., and Webster, C., "Biomarkers and surrogate end points for fit-for-purpose development and regulatory evaluation of new drugs," *Clinical Pharmacology & Therapeutics*, vol. 81, no. 1, pp. 104–107, 2007.

[113] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V., "Feature selection for SVMs," *Adv Neural Info Proc Sys (NIPS01)*, pp. 668–674, 2001.

[114] Williams, T., Toups, K., Saggese, D., Kalli, K., Cliby, W., and Muddiman, D., "Epithelial ovarian cancer: disease etiology, treatment, detection, and investigational gene, metabolite, and protein biomarkers," *J Proteome Res*, vol. 6, no. 8, pp. 2936–2962, 2007.

[115] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H., "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

[116] Wulfkuhle, J., McLean, K., Paweletz, C., Sgroi, D., Trock, B., Steeg, P., and Petricoin III, E., "New approaches to proteomic analysis of breast cancer," *Proteomics*, vol. 1, no. 10, 2001.

[117] Xu, Y., Shen, Z., Wiper, D., Wu, M., Morton, R., Elson, P., Kennedy, A., Belinson, J., Markman, M., and Casey, G., "Lysophosphatidic acid as a potential biomarker for ovarian and other gynecologic cancers," *J Am Med Assoc*, vol. 280, no. 8, pp. 719–723, 1998.

[118] Yue, Y., Finley, T., Radlinski, F., and Joachims, T., "A support vector method for optimizing average precision," in *Proceedings of the 30th annual international ACM SIGIR conference*, p. 278, 2007.

[119] Yuille, A. and Rangarajan, A., "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[120] Zhang, H., Ahn, J., Lin, X., and Park, C., "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, vol. 22, no. 1, p. 88, 2005.

[121] Zhang, Y., Kolinski, A., and Skolnick, J., "Touchstone II: A new approach to ab initio protein structure prediction," *Biophysical Journal*, vol. 85, pp. 1145–1164, 2003.

[122] Zhang, Y. and Skolnick, J., "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, no. 4, pp. 702–710, 2004.

[123] Zhang, Y. and Skolnick, J., "TASSER: An automated method for the prediction of protein tertiary structures in CASP6," *Proteins*, vol. 61(S7), pp. 91–98, 2005.

[124] Zhang, Z., Kwok, J., and Yeung, D., "Surrogate maximization/minimization algorithms and extensions," *Machine Learning*, vol. 69, no. 1, pp. 1–33, 2007.

[125] Zhou, H., Pandit, S., Lee, S., Borreguero, J., Chen, H., Wroblewska, L., and Skolnick, J., "Analysis of TASSER-based CASP7 protein structure prediction results," *Proteins: Structure, Function, and Bioinformatics*, vol. 69, no. S8, pp. 90–97, 2007.

[126] Zhou, M., Guan, W., Walker, L., Mezencev, R., Benigno, B., Gray, A., Fernández, F., and McDonald, J., "Rapid Mass Spectrometric Metabolic Profiling of Blood Sera Detects Ovarian Cancer with High Accuracy," *Cancer Epidemiology Biomarkers & Prevention*, vol. 19, no. 9, p. 2262, 2010. http://web.chemistry.gatech.edu/∼fernandez/DART_dataset.mat, http://web.chemistry.gatech.edu/∼fernandez/suppl-X1.xls.

[127] Zhou, M., McDonald, J., and Fernández, F., "Optimization of a direct analysis in real time/time-of-flight mass spectrometry method for rapid serum metabolomic fingerprinting," *Journal of the American Society for Mass Spectrometry*, vol. 21, no. 1, pp. 68–75, 2010.

[128] Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R., "1-norm support vector machines," in *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, pp. 49–56, 2004.

[129] Zou, H. and Li, R., "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, vol. 36, no. 4, p. 1509, 2008.